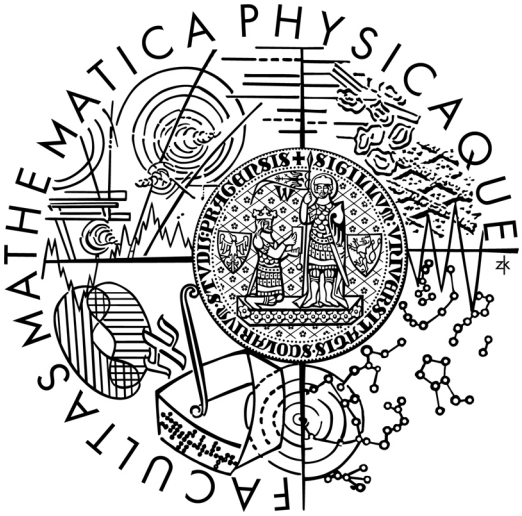
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Radka Hladíková

Data Profiling

Katedra softwarového inženýrství (KSI)
Vedoucí diplomové práce: Ing. Vladimír Kyjonka
Studijní program: Informatika, Diskrétní matematika a optimalizace

Poděkování
Na tomto místě bych ráda poděkovala vedoucímu mé diplomové práce Ing. Vladimírovi Kyjonkovi za cenné rady a také firmě SAS za licencování potřebného SW

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 14.4.2010

Radka Hladíková

Obsah

1 Úvod 6
1.1 Úvod do problematiky 6
1.2 Cíl práce 6
1.3 Obsah práce 6
2 Kvalita dat 8
2.1 Východiska 8
2.1.1 Problém kvality dat 8
2.1.2 Příčiny nekvality dat 8
2.1.3 Důsledky špatné kvality dat 8
2.2 Důležité poznámky 9
2.2.1 Posouzení správnosti dat 9
2.2.2 Relativita kvality dat 9
2.3 Nejčastější zdroje problémů v datech 9
2.3.1 Procesy vzniku dat 10
2.3.2 Procesy transformace dat 10
2.3.3 Procesy čištění dat 11
2.3.4 Časová závislost dat 11
2.4 Nejčastější problémy s kvalitou dat 11
2.5 Cíle projektů kvality dat 11
3 Řešení datové kvality 12
3.1 Aktivity v procesu kvality dat 12
3.2 Proces řízení kvality dat 13
3.2.1 Porozumění datům 13
3.2.2 Zlepšování kvality dat 14
3.2.3 “Provoz“ – rutinní udržování datové kvality 17
3.3 Metody 18
3.3.1 Softwarové nástroje 18
4 Data Profilig 21
4.1 Cíle data profilingu 22
4.2 Zisky a ztráty data profilingu 22
4.3 Popis profilingu 22
4.4 Úrovně data profilingu 22
4.5 Techniky data profilingu 23
4.5.1 Profiling atributů 23
4.5.2 Profiling vztahů 23
4.5.3 Profiling přechodu stavů 23
4.5.4 Profiling závislostí 23
4.5.5 Data gazing 23
4.6 Metriky 24
4.6.1 Definice metrik dat 24
4.6.2 Členění metrik kvality dat 25
4.6.3 Dimenze datové kvality 25
4.6.4 Příklady dimenzí datové kvality 26
4.6.5 Agregace metrik 27
5 Návrh základních metrik a metrik pro sledování defektů a jejich vyhodnocení 29
5.1 Popis problému 29

5.2 Zdrojová data ... 29
5.3 Analýza problému ... 30
5.4 Navrhovaná architektura řešení ... 30
5.5 Základní profiling ... 31
5.6 Rozšířený profiling ... 31
5.6.1 Návrh metrik pro úplnost, správnost a formát dat - dle atributů ... 31
Ukazatel plat ... 32
5.6.2 Konzistence dat ... 32
5.7 Obohacování dat ... 32
5.8 Deduplikace dat ... 33
5.9 Vyhodnocení kvality dat ... 33
5.10 Celkové vyhodnocení procesu kvality dat ... 34
6 Realizace navrženého systému vyhodnocení kvality dat ... 39
6.1 Základní profiling ... 39
6.1.1 Základní charakteristiky dat ... 39
6.1.2 Základní profiling platu ... 40
6.2 Kvalita dat ... 43
6.2.1 Ukazatel plat ... 43
6.2.2 Atribut příjmení ... 43
6.2.3 Atribut jméno ... 46
6.2.4 Atribut titul ... 47
6.2.5 Atribut pozice ... 49
6.2.6 Atribut email ... 49
6.2.7 Atribut název společnosti ... 51
6.2.8 Atribut telefon ... 52
6.2.9 Atribut město ... 54
6.2.10 Konzistence dat ... 55
6.3 Deduplikace ... 58
6.4 Obohacení dat ... 59
6.5 Stanovení hodnot výsledných metrik ... 62
6.5.1 Vyhodnocení oblasti Kvalita dat ... 62
6.5.2 Vyhodnocení oblasti Deduplikace ... 64
6.5.3 Vyhodnocení oblasti Obohacení dat ... 64
6.5.4 Výsledné vyhodnocení ... 65
7 Závěr ... 66
8 Seznam použité literatury a internetových zdrojů ... 67
9 Terminologický slovník ... 68
10 Příloha - přehled tabulek ... 69
11 Příloha - přehled souborů na CD ... 70

Název práce: *Data Profiling*
Autor: *Radka Hladíková*
Katedra (ústav): *Katedra softwarového inženýrství (KSI)*
Vedoucí diplomové práce: *Ing. Vladimír Kyjonka*
e-mail vedoucího: *Vladimir.Kyjonka@cze.sas.com*

Abstrakt:

Diplomová práce se zabývá problematikou datové kvality a data profilingem. Práce analyzuje a shrnuje problematiku datové kvality, datových defektů, procesu datové kvality, měření kvality dat a data profilingu. Hlavním tématem je data profilig jako proces zkoumání dat dostupných v existujících zdrojích dat a vytváření statistik a informací o těchto datech. Je zde navrhnut systém pro vyhodnocování stavu dat z hlediska jejich kvality. Práce se zaměřuje na měření obecných charakteristik dat, sledování datových defektů a jejich analýzu. Pro reálná data je navrhnut a za pomocí SW datové kvality realizován systém pro vyhodnocení datové kvality.

Klíčová slova: datová kvalita, data profing, metrika datové kvality, DataFlux

Title: *Data Profiling*
Author: *Radka Hladíková*
Department: *Department of Software Engineering*
Supervisor: *Ing. Vladimír Kyjonka*
Supervisor's e-mail address: *Vladimir.Kyjonka@cze.sas.com*

Abstract:

This thesis puts mind on problems with data quality and data profiling. This Work analyses and summarizes problems of data quality, data defects, process of data quality, data quality assessment and data profiling. The main topic is data profiling as a process of researching data available in existing data sources and creating of statistics and information about these data. There is a projected system for evaluating data status in term of its quality. Work is focused on measuring of general characteristic of data, following data defects and its analyses. With the help of data quality SW there is a projected and realized system for evaluation of data quality for real data.

Keywords: data quality, data profiling, data quality metric, DataFlux

1 Úvod

1.1 Úvod do problematiky

V posledních desetiletích se svět změnil z čistě průmyslové ekonomiky v ekonomiku informační. Informace se stala nejdůležitější konkurenční výhodou. To je také důvodem, proč firmy dělají maximum proto, aby získaly a vhodně využily co možná nejvíce informací. Nejde jenom o co největší prodej výrobků, ale také o to, získat co nejvíce informací.

Data a informace jsou velice důležité pro tvorbu strategických plánů společnosti a určují to, jak bude daná společnost úspěšná. Právě proto je vysoká kvalita informací pro každou společnost nesmírně důležitá. Špatná kvalita informací může mít negativní vliv na prosazení společnosti v konkurenčním boji. Nekvalitní data jsou nedostatečným vstupem pro managerské rozhodování a vedou k špatným rozhodnutím. Cílem projektů kvality dat, by tedy mělo být systematické zlepšování kvality dat na takovou úroveň, aby neovlivňovala negativně business procesy dané společnosti nebo aby naopak umožnila lepší fungování těchto procesů.

Protože je datová kvalita pojem relativní a subjektivní, je třeba mít k dispozici nějaký systém pomocí něhož budeme datovou kvalitu měřit a posuzovat.

Měření kvality dat je exaktní a statistické vyhodnocení dat, při kterém určujeme zda operační data jsou správného typu, kvality a kvantity vzhledem k jejich užití. Součástí procesu měření kvality dat je tzv. data profilig. Jedná se o proces zkoumání dat dostupných v existujících zdrojích dat a vytváření statistik a informací o těchto datech.

1.2 Cíl práce

Cílem této práce je shrnout problematiku datové kvality a problematiku měření kvality dat a navrhnout systém pro vyhodnocování stavu dat z hlediska jejich kvality, tzv. data profiling. Práce se zaměřuje především na měření a obecné charakteristiky dat a sledování a analýzu datových defektů.

1.3 Obsah práce

Následující kapitola se věnuje problému kvality dat, zabývá se důvody, proč jsou data nekvalitní a důsledky těchto nekvalitních dat. Jsou zde rozebrány nejčastější zdroje problémů v datech a diskutovány nejfrekventovanější problémy s kvalitou dat.
Třetí kapitola podrobněji popisuje samotný proces kvality dat a jednotlivé části procesu kvality dat, jako je porozumění datům, zlepšování kvality dat a rutinní provozování datové kvality. Jsou zde zmíněny nejvýznamnější dodavatelé Softwarových nástrojů pro implementaci programu kvality dat.
Čtvrtá kapitola se věnuje samotnému data profilingu. Vysvětluje pojem data profiling a rozebírá různé techniky data profilingu, jako jsou profiling atributů, profiling vztahů a profiling závislostí. Je zde vysvětlen pojem metrika, členění metrik a jejich agregace.
Obsahem páté kapitoly je návrh metrik pro posouzení datové kvality a systému pro vyhodnocení kvality dat pro reálná data. Jsou navrženy jednotlivé metriky kvality dat - úplnost, správnost a konzistence dat, metrika určující úspěšnost deduplikace dat a metrika

posuzující úspěšnost obohacení dat. Na základě těchto dílčích metrik je navrhnut systém posouzení měření kvality dat
Šestá kapitola se věnuje implementaci navrženého systému vyhodnocení dat.
Sedmá kapitola obsahuje stručné shrnutí výsledků práce.
V osmé kapitole je zmíněna použitá literatura a použité internetové zdroje.
Devátá kapitola obsahuje terminologický slovník.
Desátá kapitola obsahuje seznam tabulek.
V jedenácté kapitole jsou vyjmenovány a stručně popsány adresáře a soubory uložené na přiloženém CD.

2 Kvalita dat

Tato kapitola se věnuje obecné problematice kvality dat, shrnuje důvody, proč jsou data nekvalitní a rozebírá důsledky těchto nekvalitních dat. Věnuje se nejčastějším zdrojům problémů v datech a diskutuje nejfrekventovanější problémy související s kvalitou dat a popisuje cíle projektů kvality dat.

2.1 Východiska

2.1.1 Problém kvality dat

S problémem datové kvality se setkáváme ve všech společnostech bez ohledu na kvalitu jejich organizace. V každém systému jsou chyby v datech a snahou by mělo být, aby jich tam bylo co nejméně. V dnešní době se s daty často pracuje automatizovaně a špatná kvalita značně snižuje kvalitu a použitelnost tohohle automatického zpracování. Výsledky automatického zpracování dat jsou často nesprávné a to následně vede k nesprávnému fungování podnikových procesů a ekonomickým ztrátám.

2.1.2 Příčiny nekvality dat

Existují dvě hlavní příčiny toho, proč jsou data nekvalitní. Prvním důvodem je to, že data do informačních systémů vkládají lidé a lidé dělají chyby, ať už jsou to překlepy, zápisy do nesprávných polí, neznalost, špatné porozumění a nebo různá „lidská tvořivost“.
Druhým důvodem je nekonzistence informačních systémů podniku. Informační systémy se skládají z mnoha různých systémů založených na různých technologiích, tyto systémy implementují různí dodavatelé. Postupně se zavádějí nové systémy, často bez ohledu na ty stávající a každý systém je často orientován produktově a bez ohledu na ostatní systémy společnosti.

2.1.3 Důsledky špatné kvality dat

Nekvalitní data problémem v IT i businessu

Chyby v datech způsobují společnosti problémy jak v IT, tak v businessu. Z pohledu IT znamená špatná kvalita dat vyšší časové a finanční nároky a nákladnější a náročnější realizaci projektů.
Ze strany businessu má špatná datová kvalita vliv na úspěch společnosti na trhu. Vede například k zvýšení nákladů na různé procesy, nespokojenosti zákazníků či k ztrátě důvěryhodnosti firmy.

Problém integrace dat

Chyby v datech jsou problémem také při procesu integrace dat. Proces integrace dat je sjednocení dat z různých částí informačního systému společnosti, propojení jednotlivých informačních systémů a jejich ztotožnění s fungováním business procesů. Cílem integrace je vyšší efektivita, zrychlení a zpřehlednění těchto systémů a tedy v důsledku zlevnění jejich realizace a vytvoření konkurenční výhody. V praxi se setkáváme s různými přístupy k integraci dat, které se liší různými parametry. Jedná se například o datové sklady a business inteligenci, podnikovou aplikační integraci, datovou integraci, integraci podnikových procesů, podnikovou aplikační integraci. Všechny tyto přístupy mají ale společné to, že zajišťují automatizované předávání a sdílení dat mezi všemi částmi podnikových systémů. Pokud mají tedy data v některé části tohoto systému špatnou kvalitu, chyby v těchto datech se

automaticky šíří i do ostatních částí systému. V případě, že data v jednotlivých částech systému jsou vzájemně nekonzistentní, integrace dat nemůže fungovat správně.

Nutnost specializovaného SW pro kvalitu dat

Většina softwarových produktů zaměřená na hromadné zpracování dat se v jisté míře snaží vypořádat s nekvalitními daty. Tyto produkty však často disponují z pohledu datové kvality jen omezenou funkčností, často nejsou znovupoužitelné a nemají možnost zobecnění.Ukázalo se že použití nespecializovaného software je nejen příliš pracné ale i málo účinné a příliš se tedy neosvědčilo.
Datová kvalita je komplexní a složitý problém a proto byly vyvinuty specializované komerční produkty, které slouží především k analýze a čištění dat.

2.2 Důležité poznámky

2.2.1 Posouzení správnosti dat

V ideálním případě bychom byli bez problému schopni zjistit, zda jsou data společnosti odpovídající kvality a pokud ne, byli bychom schopni chyby nalézt a opravit. V praxi bývá ale situace mnohem komplikovanější. Jediným způsobem, jak ověřit kvalitu dat, je porovnání s nějakým důvěryhodným zdrojem dat, který je v každém okamžiku správný. Bohužel takový zdroj ve většině případů buď neexistuje nebo není snadno dostupný.

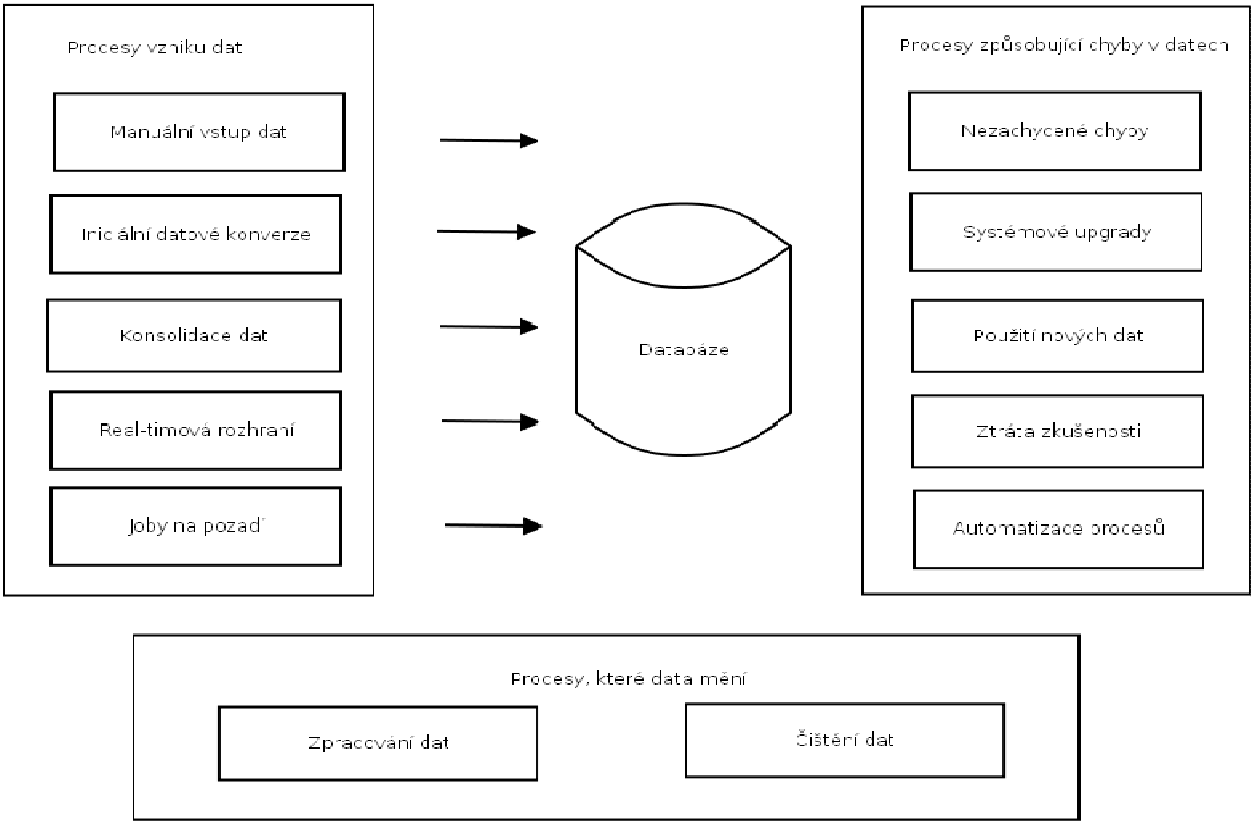
2.2.2 Relativita kvality dat

Datová kvalita je hodně relativní a subjektivní. Proto je obtížné stanovit, co je dostatečná kvalita dat. Tahle úloha je důležitou součástí jakékoliv analýzy datové kvality a základním východiskem pro jakoukoliv další aktivitu v oblasti kvality dat.
Je důležité si uvědomit, že není v lidských silách dosáhnout stoprocentní kvality dat.
V ideálním případě dosáhneme co největší kvality dat vzhledem k vynaloženým prostředkům.

2.3 Nejčastější zdroje problémů v datech

Data jsou zpracovávána velkým počtem různých procesů, kde každý proces svým způsobem ovlivňuje kvalitu dat.

Na obrázku *Procesy ovlivňující kvalitu dat* jsou zobrazeny procesy, které mají vliv na kvalitu dat.



Procesy ovlivňující kvalitu dat

2.3.1 Procesy vzniku dat

První skupinou procesů ovlivňující kvalitu dat, jsou procesy, pomocí kterých se data do databáze dostávají (zobrazeny v levé horní části obrázku).
Mezi takové procesy patří např. manuální vstup dat, kdy jsou data do databáze vkládána přímou akcí uživatele. Další možností je automatizovaný vstup dat (ať už se jedná o proces konsolidace dat nebo strohé zkopírování dat z jiného zdroje). V tomto případě se data do databáze dostávají přes různá real-timeová rozhraní případně pomocí jobů na pozadí přenášejících data v různých časových intervalech. Spousta takto získaných dat může být chybná už ve zdroji a je jen jednoduše převedena do cílové databáze.
Obvykle se na vstupujících datech provádí tzv. iniciální datové konverze, např. úpravy dat do nějakého specifického formátu, tyto konverze také bývají častým zdrojem chyb.

2.3.2 Procesy transformace dat

Druhou skupinou procesů způsobujících změnu kvality dat jsou procesy, které data v databázi transformují, nebo s nimi pracují (zobrazeny v pravé horní části obrázku). Možným zdrojem problémů je například automatizace procesů a nezachycené chyby v datech, které se v průběhu automatizovaných datových transformací šíří dále. Problémy mohou způsobit také upgrady systémů na vyšší verze či redesign databáze. Z uživatelského pohledu bývá

problémem nové použití dat, se kterým nejsou uživatelé dostatečně seznámeni, případně ztráta zkušenosti s dosavadním používáním dat (například odchodem klíčového zaměstnance).

2.3.3 Procesy čištění dat

Třetí skupinou procesů ovlivňujících kvalitu dat, jsou procesy, kterými se datová kvalita zvyšuje (v dolní části obrázku). Datová kvalita je zpravidla pozitivně ovlivňována čištěním dat a následným zpracováním závislých dat.

2.3.4 Časová závislost dat

Negativní vliv na datovou kvalitu mají také procesy, které zobrazují sice správná data ale v nesprávný čas. Data sice nebyla modifikována, ale nejsou aktuální a tedy správná. Tahle situace zpravidla nastává, pokud se nějaká skutečnost změní, ale tahle změna není do databáze zavedena.

2.4 Nejčastější problémy s kvalitou dat

Mezi nejčastější problémy s kvalitou dat patří :

- Nestrukturovaný zápis údajů
- Nesprávné použití diakritiky, případně chybějící diakritika
- Nesprávná interpretace NULL hodnot, například různé použití uměle vytvořených NULL hodnot pro datumová pole (1.1.1910)
- Duplicity
- Neúplné záznamy
- Neaktuální data
- Implicitní předem vyplněné hodnoty, které uživatel bezmyšlenkovitě potvrdí
- Přepsání či přeslechnutí při vkládání dat
- Různá pravidla v různých systémech, např. různá pravidla na formát telefonního čísla
- Povinné položky, u kterých není zřejmé co do nich má uživatel vyplnit
- Nedostatečné kontroly vstupních dat
- Ztráta některých dat
- Nedostupnost dat

2.5 Cíle projektů kvality dat

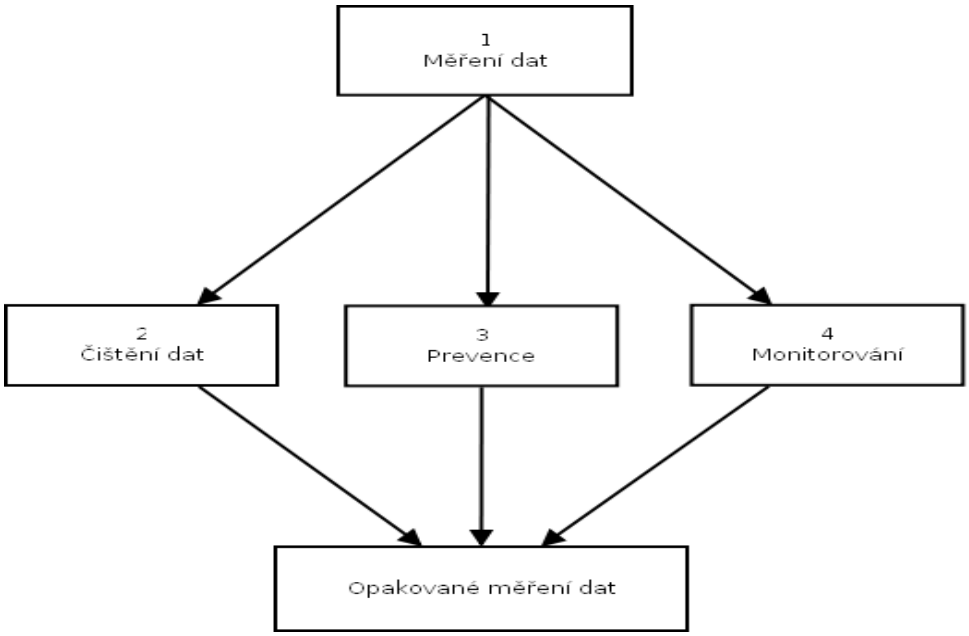
Cílem projektů kvality dat, by mělo být zlepšení kvality dat na takovou úroveň, aby neovlivňovala negativně business procesy dané společnosti, nebo aby naopak umožnila lepší fungování těchto procesů. Zlepšování kvality dat není jednorázovou akcí, ale je během na dlouhou trať. Jedinou možností jak zaručit kvalitu dat je systematický program, který měří a zlepšuje kvalitu dat, využívá nepřetržitého monitorování datové kvality a vytváří preventivní opatření, která zabrání možným problémům s datovou kvalitou. Jak už jsem zmínila dříve je datová kvalita pojem relativní a subjektivní, proto je třeba mít k dispozici sadu metrik, pomocí nichž budeme datovou kvalitu posuzovat.

3 Řešení datové kvality

V následující kapitole popíšeme, jak řešit datovou kvalitu, postupně se budeme věnovat měření kvality dat, čištěním dat, monitorováním dat a jednotlivým preventivním opatřením v programu kvality dat.

3.1 Aktivity v procesu kvality dat

Komplexní program zabývající se kvalitou dat se skládá z několika různých procesů, které jsou podrobněji zobrazeny na schématu *Aktivity v procesu kvality dat.*



Aktivity v procesu kvality dat

Prvním krokem je změřit kvalitu dat. Zjišťování kvality dat označujeme termínem data quality assessment, neboli **měření kvality dat** (viz 1 na schématu). Cílem této aktivity je identifikace chyb v datech a měření jejich vlivu na business procesy dané společnosti. Poté co dokončíme proces měření kvality dat, jsme schopni rozhodnout o kvalitě dat.

Většinou je třeba nejprve zlepšit kvalitu již existujících dat, tento proces označujeme pojmem data cleasing, neboli procesem **čištění dat** (viz 2 na schématu). Cílem čištění dat je opravit většinu existujících chyb v datech a udělat z chybných dat data, kterým lze důvěřovat, a která lze bez obav využívat.

Poté co dosáhneme požadované kvality dat, tj. takové kvality, která odpovídá business potřebám dané společnosti, musí být dalším krokem zabránění zavlečení nových chyb dat do systémů pomocí různých **preventivních opatření** (viz 3 na schématu). Tento proces bývá zpravidla mnohem komplikovanější než identifikace chyb v již stávajících datech a proces jejich odstranění.

Dalším neméně důležitým procesem v programu kvality dat je tzv. ongoing monitoring, neboli **nepřetržité monitorování** (viz 3 na schématu). Bez nepřetržitého monitorování se může stát že, i když budou vybudována ta nejlepší preventivní opatření, že si případných chyb v datech všimneme příliš pozdě.

Je třeba **opakovaně měřit kvalitu dat**, porovnávat výsledky měření a včas si všímat změn v datové kvalitě. Výsledkem takovéto měření musí být posléze další čištění dat a tvorba nových preventivních opatření.

3.2 Proces řízení kvality dat

Proces řízení kvality dat se skládá z několika částí. Z části poznání a porozumění datům, z procesu zlepšování kvality dat a z procesu monitorování dat a reportingu.

3.2.1 Porozumění datům

Porozumění datům zahrnuje aktivity jako je analýza business procesů a pravidel v dané společnosti, dále také měření dat (data assessment) a analýza dat (data profiling), porozumění sémantice dat, analýze chyb v datech.
Měření kvality dat je základním prvkem každého procesu řízení kvality dat. Proces měření kvality dat slouží k identifikaci chybných dat a zjištění jejich vlivu na různé business procesy společnosti. Jednou z možností usnadňujících zjišťování kvality dat jsou moderní databáze, které nám umožňují hromadné a rychlé zpracování dat a umožňují nám také např. automatickou validaci dat.

Pravidla kvality dat

Ověřování kvality dat je založeno na **pravidlech kvality dat** (tzv. data quality rule). Jedná se o omezení, která validují datový element nebo vztah mezi několika datovými elementy a mohou být implementována jako počítačové programy. Termín vztah mezi daty tu chápeme v širším smyslu – od nejjednodušších vztahů až po komplikovaná business pravidla. Řešení spočívá v návrhu a implementaci stovek až tisíců takovýchto pravidel a jejich použití při hledání nekonzistencí.
Je samozřejmě velice obtížné, ba téměř nemožné, navrhnout perfektní pravidla datové kvality. Velkým problémem návrhu bývá to, že některá pravidla selhávají v detekci chyb a naopak označí správná data za data chybná.

Reporty kvality dat

Pomocí pravidel datové kvality a specializovaných nástrojů získáváme **reporty zobrazující chyby v datech**. Každá chyba odpovídá jednomu nebo několika datovým elementům z jedné nebo několika tabulek. Pochopení těchto reportů a jejich využití bývá složité. Ověřování kvality dat používá a vytvoří mnoho dalších typů metadat - jako jsou datové modely, datové katalogy, datové profily, definice pravidel, agregované metriky dat. Organizace těchto metadat do využitelného datového skladu s zabudovanou dimenzionální ohodnocovací tabulkou dimenzionálních dat bývá dalším úkolem v programu kvality dat.

Aplikace výsledků měření dat

V níže uvedeném seznamu je krátký seznam aplikace výsledků úspěšného měření kvality dat:

1) Pomáhá popsat stav dat, porozumět jak data podporují procesy a odhadnout vliv datových problémů na procesy společnosti

2) Pomáhá plánovat procesy čištění dat a vyhodnocovat výsledky čištění dat

3) Zjednodušuje procesy konverzí a konsolidací dat poskytováním informací o kvalitě dat ve zdrojovém systému

4) Pomáhá porozumět zdrojům existujících datových problémů a pomáhá navrhnout způsoby zlepšení procesů

5) Pomáhá porozumět implikacím mezi nově plánovaným používáním dat a procesy, předtím, než budou data použita

6) Pomáhá při testování upgradu systému, případně při velkých změnách systému tím, že za testovací případy mohou být vybrány datové záznamy s chybami. A to pomáhá předpovědět jak budou změny pracovat s reálnými ne vždy perfektními daty.

3.2.2 Zlepšování kvality dat

Data téměř každé společnosti jsou plné chyb a díky těmto chybám není možné data efektivně a rychle v daných situacích využívat, případně může použití chybných dat vést k špatným managerským rozhodnutím a ztrátám. Abychom mohli data maximálně využít, je třeba tyto chyby v datech odstranit.
Proces opravy chybných datových elementů nazýváme **čištění dat**. Přestože chyby dat jsou obvykle rozšířeny do všech částí databáze, čištění dat se většinou zaměřuje na standardizaci zákaznických dat, odstranění duplicit a matching. Čištění ostatních dat většinou zůstává většinou v pozadí.

Čištění dat

Aby bylo čištění dat úspěšné, je nezbytné, aby navázalo na komplexní proces měření kvality dat. Měření musí probíhat také pravidelně v průběhu čištění dat a i poté co je dokončeno, abychom identifikovali nové problémy s daty a ověřili, že se kvalita dat zlepšila.

V procesu čištění dat využíváme a kombinujeme dvě metody. Jednou je manuální čištění dat lidskou silou a druhou je odstraňování dat automaticky, tj. počítačovým programem. Obě tyto metody se vzájemně doplňují.

Manuální a automatické čištění dat

Manuální čištění dat spočívá v ručním odstraňování chyb v datech, záznam po záznamu. Využívá se zejména tam, kde chyby v záznamech jsou příliš odlišné a nelze najít nějaký vztah mezi nimi. Problémem manuálního čištění dat je obvykle velká časová a finanční náročnost tohoto řešení.

Většina chyb v datech však není úplně libovolných, ale je vytvořena nějakým systematickým procesem. A pokud je nějakým takovým procesem vytvořeno větší množství chyb, je možno vysledovat vzor tvorby těchto chyb. Poté je možné udělat automatické opravy všech podobných chyb jednotným čistícím algoritmem. Takový proces čištění dat označujeme jako **automatické čištění dat**.

Problém závislosti pravidel datové kvality

Jedním z problémů čištění dat založeného na pravidlech je to, že pravidla datové kvality jsou vzájemně závislá. Oprava datového elementu porušujícího jedno pravidlo často vede k narušení pravidla jiného. Může se stát, že proces čištění nám tedy do systému zavede spoustu nových chyb i přestože stávající opraví.

Automatické čištění dat a návrat k původním datům

Dalším problémem vyvstávajícím v procesu čištění dat je to, že automatizovaný proces čištění dat spočívá v počítačových programech, které dělají opravy hromadně. Existuje tedy riziko, že opomeneme některé výjimky v logice, nebo že v programu bude chyba. Abychom si poradili s touto situací je nezbytné vytvořit pro tento účel nějaký auditní systém. Tento mechanismus bude zachycovat všechny změny provedené v datech a bude schopen jednoduše vrátit stav dat do původního stavu, dokud algoritmus čištění neposkytne akceptovatelné výsledky.

Automatické čištění dat

Cílem automatických oprav je identifikace datových defektů a jejich následná automatická oprava založená na pravidlech datové kvality.

Proces automatického čištění dat se obvykle skládá z několika kroků :

Parsing

Cílem parsingu je rozeznat sémantické prvky uložené v atributu a vyprodukovat z nich jednotlivé datové komponenty. Parsování záznamů používáme v případě, že zkoumaná množina dat obsahuje v jednom datovém poli více sémantických komponent např. jméno, příjmení, adresu, město stát, název společnosti. Cílem parsování tohoto záznamu je potom rozdělit informace z datového pole do jednotlivých datových komponent.

Abychom toho dosáhli je třeba mít k dispozici:

- Jména a typy datových komponent (metadata), které očekáváme, že nalezneme v parsovaném záznamu
- Množinu validních hodnot (domény) pro každý typ datové komponenty
- Přípustné formy, které mohou data mít – např. jejich syntaxe, počet elementů
- Prostředky pro označení záznamů, která mají neidentifikovaná data

Standardizace

Standardizace je proces transformace datových elementů do standardního formátu datového elementu, tj. každý prvek nějaké skupiny datových elementů musí odpovídat danému formátu. Standardním formátem rozumíme takovou reprezentaci hodnot, jež je pevně daná pravidly. Standardní formát často využíváme v případě konsolidace, kdy se snažíme propojit jednotlivé entity dohromady. Častým příkladem standardizace je standardizace křestních jmen, kdy se zkráceniny a přezdívky křestního jména transformují do standardního tvaru, často se také setkáváme se standardizací rodných či telefonních čísel.

Unifikace

Pod pojmem unifikace rozumíme identifikaci záznamů a seskupení těchto záznamů do skupin, které patří ke konkrétnímu subjektu (např. k osobě, firmě). Dále se provede výběr nejlepšího záznamu, tzv. master záznamu a skupině záznamů se přidělí (nový) jednoznačný identifikátor. V procesu unifikace se využívají pravidla, která jsou řízeny business pravidly dané společnosti. Tato pravidla s ohledem na shodu nebo alespoň podobnost v jednotlivých atributech definují, které záznamy s větší nebo menší pravděpodobností patří k sobě.

V procesu unifikace bývají rozhodující následující kritéria :

1) Priorita systému z nějž data pochází
2) Datum pořízení datového záznamu
3) Hodnoty atributů datového záznamu

Eliminace duplicit (Deduplikace)

Eliminace duplicit je proces hledání různých reprezentací jedné a tytéž entity v množině dat a eliminace všech těchto reprezentací mimo jedné určené.
Deduplikované databáze tedy obsahují právě jeden záznam pro každého konkrétního jedince – tzv. representant. Po procesu deduplikace by mělo být tedy záznamů stejně nebo méně než před ním.
V některých případech, např. u primárních klíčů tabulek v relační databázi nejsou duplicity povoleny a proto je nutné nalézt duplicitní záznamy a redukovat je na jednu entitu.

Seskupení dat (Merge / Purge)

Seskupení dat je podobné jako eliminace duplicit, s tou výjimkou, že zatímco eliminace duplicit pouze odstraňuje zdvojené záznamy tak seskupování dat spočívá v agregaci mnohonásobných datových entit a následné eliminací duplicit.

Rozšířená datová kvalita

Obohacování dat

Proces obohacování dat je proces zvyšování informační hodnoty existujících dat přidáním chybějící nebo nedostupné informace z jiného externího zdroje, jako je například obchodní rejstřík, seznam poštovních směrovacích čísel apod..

Interní

Interním obohacováním rozumíme obohacovaní ze samotného zdroje dat, bez využití zdroje jiného. Příkladem je doplnění pohlaví, případně data narození z rodného čísla osoby.

Externí

Pod pojmem externí obohacování dat rozumíme doplňování informací z externích zdrojů. Příkladem může být doplnění dat z údajů obchodního rejstříku, z katastru nemovitostí, doplňování poštovních údajů apod. Často se také využívá tzv. geocoding. Geocodingem rozumíme doplnění údajů o zeměpisné šířce a délce k danému adresnému bodu.

Householding

Householding je proces spojování záznamů, které mají něco společného, do jedné množiny. Slovo „household“ chápeme v tomto kontextu v přeneseném významu. Může se jednat o domácnost, rodinu, zájmovou skupinu, profesní skupinu ale také komerční subjekty, které mezi sebou mají např. vlastnické nebo partnerské vztahy, nebo se může jednat o skupinu zaměstnanců, vlastníků atd.

Typy householdingu :

Osobní householding

Osobní householding je takový householding, kde seskupování záznamů je prováděno na základě vztahů mezi konkrétními fyzickými osobami.
Hlavní využití osobního householdingu spočívá v úspoře nákladů při oslovování pouze jednoho člena rodiny zákazníků, v profitabilitě zákazníka, kdy se firmy zaměří zejména na zákazníky, kteří generují zisk, v křížovém prodeji – prodej založený na principu, že o některých významných výdajích rozhodují členové domácnosti dohromady.

Komerční householding

Komerčním householdingem rozumíme takové seskupování záznamů, kde za jednu skupinu bereme obchodní jednotky.
Komerční householding se nejčastěji využívá k rozpoznání finančních nebezpečí (odhalení úvěrových rizik, rozpoznání rizik bankrotu, rizika spjatá s národností dané firmy,..), ke konsolidaci účetnictví a k řízení různých obchodních operací

Kombinace osobního a komerčního householdingu

Dalším typem householdingu je kombinace osobního a komerčního householdingu. V případě kombinovaného householdingu jde o seskupování založeném na vztahu fyzické osoby a obchodních subjektů.
Kombinovaný householding se využívá např. k marketingovým účelům, k odhalení podvodného jednání, k identifikaci různých střetů zájmů apod.

3.2.3 “Provoz“ – rutinní udržování datové kvality

Rutinní provozování

Ve fázi rutinního provozování pravidelně kontrolujeme kvalitu dat a identifikujeme potenciální zdroje problémů a poskytujeme odezvu, zda byly předchozí akce zlepšující kvalitu dat úspěšné či nikoli. Podle výsledků jsou potom implementovány nové akce, aby bylo docíleno co možná největší datové kvality.

Monitorování a reporting

V procesu monitorování a reportingu dochází k nepřetržitému monitorování dat, v této fázi také vznikají nové metriky dat, důležité pro měření datové kvality. Reporting nám poskytuje zejména popis stavu vstupních dat, zobrazuje provedené změny a náhrady, četnost a typy chyb datech . Poskytuje podklady pro rozhodování při poloautomatickém zpracování a podklady pro ruční opravy dat.

3.3 Metody

Datová kvalita je komplexní problém se kterým se setkáváme ve většině společností a není snadné ho řešit. Jeho řešení spočívá v několika částech, které společně zahrnují celý problém komplexně. Tyto části mohou být využity odděleně, nicméně takové použití nepřináší očekávané výsledky, je to pouze první krok v řešení problému kvality dat.

Jedná se o následující části :

1) Softwarové nástroje pro analýzu, čištění, konsolidaci a monitoring datové kvality
2) Metodologie / Know How / Best practices – implementační tým pracuje s teoretickým a praktickým know-how vztahujícím se k managementu datové kvality a zkušenosti s implementací projektů datové kvality
3) Data Quality Management environment – změny v business procesech, organizační opatření a tvorba dokumentů, které umožní účinný management datové kvality

3.3.1 Softwarové nástroje

Funkce nabízených SW řešení

Na trhu software s datovou kvalitou je více poskytovatelů, kteří nabízejí různé softwarové produkty. Tyto produkty obvykle pokrývají hlavní požadavky systému kvality dat jako jsou:

- Data profilig
- Parsing a standardizace
- Čištění dat
- Matching
- Monitoring
- Obohacování dat

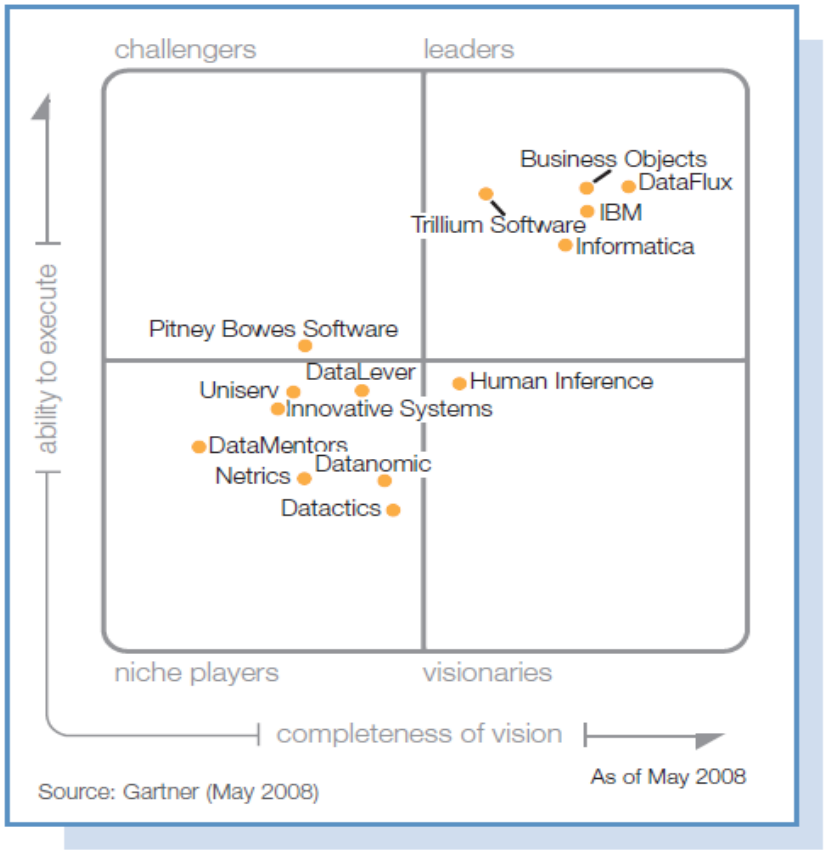
Navíc tyto produkty poskytují související funkce, které jsou potřebné pro mnoho funkcí datové kvality nebo pro specifické aplikace datové kvality.
Mezi tyto doplňující funkce patří například :

- Konektivita / adaptéry
  Softwarové nástroje mají schopnost interakce s různými typy datových struktur.
- Podpora specifických subjektů
  Softwarové nástroje disponují standardizačními schopnostmi pro specifické datové subjekty
- Mezinárodní podpora
  Softwarové nástroje podporují operace datové kvality v mezinárodním prostředí
- Metadata management
  Softwarové nástroje jsou schopny zachytit a odsouhlasit metadata vztahující se k datům
- Konfigurační prostředí
  Softwarové nástroje obsahují funkce pro vytvoření, správu a implementaci pravidel datové kvality.
- Operace a administrace
  Softwarové nástroje obsahují funkce pro podporu, správu a kontrolu datové kvality
- Servisně orientovaný SW
  Softwarové nástroje jsou servisně orientované a podporují SOA.

Nejvýznamnější softwarové produkty z oblasti datové kvality

Mezi nejvýznamnější dodavatele SW pro kvalitu dat patří zejména významní dodavatelé podnikových informačních systémů. Jedná se o firmy SAS, Informatica, IBM, SAP a další.

Na obrázku *Magic quadrant* (převzat z [5]) jsou znázorněni nejvýznamnější hráči na trhu datové kvality



Magic quadrant

DataFlux

DataFlux patří mezi nejvýznamnější produkty z oblasti nástrojů datové kvality. Nástroj se skládá z klientské aplikace dfPower Studio, sloužící zejména k tvorbě business pravidel, a DataFlux Ingegration Serveru, tedy báze integrující datovou kvalitu napříč celou organizací. Platforma DataFlux obsahuje nástroje pro data profiling, matching, čištění dat a monitorování dat. Jádro řešení DataFlux tvoří tzv. QKB (Quality Knowledge Base), obsahující ke každému sémantickému typu definice a jim odpovídající gramatiky, fonetické knihovny, knihovny regulárních výrazů, standardizační schémata a slovníky.
DataFlux poskytuje lokální podporu 36 zemím a 18 jazykům, většina zákazníků, ale využívá pouze anglické prostředí.

Informatica

Portfolio společnosti Informatica zahrnuje silnou funkcionalitu data profiligu (Data Explorer), parsingu, standardizace a matchingu (Data duality) .Naopak slabší stránkou produktu je omezená využitelnost v vícejazyčných a multicountry implementacích. Stejně tak slabší stránkou jsou operace související s validací adres geocodingem.

IBM

Hlavní platformou produktů kvality dat společnosti IBM je Information Server, ale IBM také používá měření kvality dat pomocí IBM Global Business Services Information Analyzer (nástroj na profilig a analýzu dat) a Quality Stage (parsing, standardizace a sofistikovaný matching). Nově navržený Information Server zahrnuje kromě nástrojů pro kvalitu dat také ETL nástroje, replikace a meta data management.

Trillium Software

Harte-Hanks Trilium Software poskytuje širokou škálu nástrojů datové kvality, včetně profilingu dat (TS Discovery) a vytváření dashbordů datové kvality (TS Insight). Schopnosti obohacování dat jsou zaměřené na zákaznická data (adresy, geocoding). Trilium se snaží rozšířit svou pozici a schopnosti od základní funkcionality kvality dat k tomu čemu se říká „datová inteligence“, tedy k lepšímu pochopení metadat a managementu, sémantickému pochopení a interakci business uživatelů. Reference zákazníků potvrzují spokojenost s výkonností a škálovatelností Trilium nástrojů a velice pozitivní jsou také zkušenosti se servisem a supportem.

SAP Business Objects

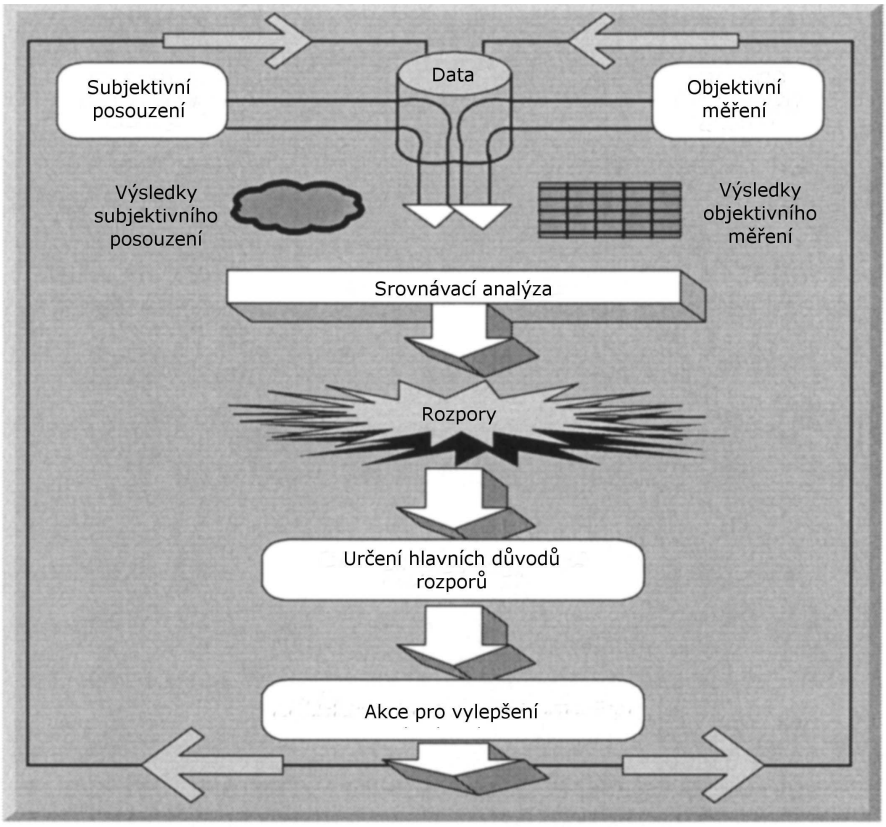
Business Objects se staly součástí portfolia společnosti SAP. Společnost SAP má širokou základnu zákazníků v oblasti ERP, u kterých je velký potenciál v oblasti datové kvality . Business Objects nabízejí širokou škálu funkcionalit datové kvality zahrnující data profilig (Data Insight XI), obvyklé operace čištění dat (Data Quality XI). Síla Business Objects je v aplikaci datové kvality, hlavně v matchingu, deduplikaci, standardizaci jmen a adres a validaci. Data profilig zůstává relativní slabinou tohoto produktu. Prostorem k zlepšení je především zvýšení spolehlivosti IQ Insight a integrace s Data Quality XI.

4 Data Profilig

Datová kvalita je velice subjektivní. Pro některou skupinu uživatelů mohou být data dostatečně kvalitní pro jinou naopak naprosto nevyhovující. Abychom byli schopni určit, jak jsou data pro dané uživatele a daný účel kvalitní, je třeba jejich kvalitu změřit.

Měření kvality dat (Data Quality Assessment) je exaktní a statistické vyhodnocení dat, při kterém určujeme, zda operační data jsou správného typu, kvality a kvantity vzhledem k jejich užití.

Obrázek *Měření kvality dat* názorně zobrazuje měření kvality dat v procesu datové kvality :



Měření kvality dat

Měření kvality dat je základním krokem v procesu kvality dat, kterým je vždy třeba začít a který je nutnou podmínkou pro kroky ostatní. Až v návaznosti na něm je možné přejít k dalším krokům jako je např. čištění dat.

Součástí procesu měření kvality dat je tzv. **data profilig**. Jedná se o proces zkoumání dat dostupných v existujících zdrojích dat a vytváření statistik a informací o těchto datech.

4.1 Cíle data profilingu

Cílem data profilingu je lépe a hlouběji porozumět stavu dat v dané chvíli a získat komplexní pohled na data společnosti. Data profiling většinou provádíme před zahájením datové integrace, migrace nebo jako součást procesu datové kvality.

4.2 Zisky a ztráty data profilingu

Díky data profilingu jsme schopni identifikovat nejdůležitější problémové oblasti v datech, jsme schopni odhadnout, jak lze nově používat data. Data profiling zlepšuje efektivitu procesu čištění dat, redukuje problémy s integrací dat, redukuje manuální analýzu dat a tedy i optimalizuje využití lidských zdrojů.

Naopak data profiling neodhalí všechny problémy datové kvality a může být tedy využit pouze jako doplněk k manuální analýze a rozhovory s business uživateli.

4.3 Popis profilingu

- Data profiling identifikuje problémy datové kvality využitím různých statistických metod
- Analyzuje a reportuje hodnoty dat, statistiky o nich, frekvence výskytu a rozsahy hodnot
- Identifikuje nekorektní datové formáty, duplikovaná data a překlepy
- Identifikuje redundantní a chybějící data, hledá a potvrzuje klíčové hodnoty a datová pravidla

4.4 Úrovně data profilingu

Data profing může být proveden v několika úrovních. Je možné dělat data profiling na datech provozních systémů a to buď přímo, a nebo abychom zabránili performance problémům, tak nad textovými soubory či databází vytvořenými z dat provozních systémů. Dále je možné provádět data profiling nad daty v datových skladech a to buď v stage, tj.datové oblasti, kam se data dostanou po nahrání do datového skladu, čili data netransformovaná a nebo přímo profiling dat v datovém modelu datového skladu. Další možností je profiling dat v Data Marts.

Levely data profilingu, tj. data profiling nad jednotlivými systémy shrnuje tabulka *Levely data profilingu*

Level	Systém	Komentář
1	**Provozní systémy**	Profiling dat přímo v provozních systémech společnosti
		Profiling dat vygenerovaných do textových souborů z operačních systému společnosti
2	**Datové sklady**	Profiling dat v staging area
		Profiling dat přímo v datovém modelu
3	**Závislé Data Marts**	Profiling dat uvnitř data marts - např. po transformacích, nebo v případě performance problémů
4	**Nezávislé Data Marts**	Profiling dat, které nejsou v datovém skladě a která nejsou přístupná z provozních systémů

Tabulka 1-Levely data profilingu

4.5 Techniky data profilingu

Mezi techniky využívané v profilingu patří dle [3] :

4.5.1 Profiling atributů

Profiling atributů spočívá ve zkoumání hodnot jednotlivých datových atributů a poskytování informací o základních agregačních statistikách (minimum, maximum), frekvenci daných hodnot a distribuci hodnot pro každý atribut.

4.5.2 Profiling vztahů

Úkolem profilingu vztahů je identifikace klíčů entit a vztahů mezi entitami. Výpočet výskytů každého vztahu v datovém modelu.

4.5.3 Profiling přechodu stavů

Profiling přechodů vztahů je kolekce technik pro analýzu životního cyklu stavově závislých objektů. Výsledkem takového profilingu je aktuální informace o pořadí a trvání jednotlivých stavů a akcí

4.5.4 Profiling závislostí

Profiling závislostí slouží k hledání skrytých vztahů mezi hodnotami jednotlivých atributů.

4.5.5 Data gazing

Data gazingem rozumíme proces hledání dat a pokus o rekonstrukci příběhu, který se za daty skrývá. Získaný reálný příběh pomáhá identifikovat parametry, o které se mohlo a nebo nemuselo jednat. Jakmile zjistíme, že příběh schovaný za jednotlivými datovými elementy nedává smysl, vytvoříme pravidla datové kvality tak, abychom zachytili data, která neodpovídají.

4.6 Metriky

Abychom mohli měřit kvalitu dat, je třeba zavést nějaká měřítka, pomocí kterých kvalitu dat kvantifikujeme, tyto měřítka označujeme jako metriky kvality dat.

4.6.1 Definice metrik dat

Metodologie návrhu metriky

Definice metriky dat musí být prováděna na základě nějaké metodologie a je třeba, aby ji následovalo několik kroků ověřujících spolehlivost navrhované metriky.
Prvním krokem definice metriky je její návrh. Při návrhu metriky je třeba vzít v úvahu specifické charakteristiky systému, jehož kvalitu měříme a zkušenosti designerů tohoto systému. Dalším krokem v definici metriky je validace této metriky. Formální validace nám pomáhá zjistit, kdy a jak metriku aplikovat.

Definiční tabulka metriky

Metriky bývají často velice komplexní a mohou být definovány pomocí mnoha atributů. Nejdůležitější atributy bývají většinou zaznamenány v tabulce definice metriky. Taková tabulka obsahuje většinou informace o metrice samotné (název metriky, účel metriky), informace vztahující se k výpočtu metriky (datový typ, jednotky, postup výpočtu) a detaily k agregaci metriky.

Příklad definiční tabulky metriky :

Metrika	**Název**	Název metriky
	ID	Jednoznačný identifikátor metriky
	Popis	Popis a účel metriky
Výpočet	**Jednotky**	Jednotky, ve kterých bude metrika vyjadřována
	Datový typ	Datový typ metriky
	Postup výpočtu	Způsob, jakým je metrika vypočítávána, včetně způsobu, jak získáme zdrojová data pro výpočet
	Rozsah hodnot	Rozsah akceptovatelných hodnot pro danou metriku
Agregace	**Nadřazená metrika**	Identifikace metriky k jejímuž výpočtu daná metrika slouží
	Podřazená metrika	Identifikace metriky z níž byla daná metrika vypočtena
	Vztah k procesu	Procesy, ve kterých bude metrika využita
	Vztah k dimenzi	Název dimenze, ke které daná metrika patří
	Agregace (Hierarchie metrik)	Metoda agregace při drill up v hierarchii metrik
	Agregace (ostatní dimenze)	Metoda agregace při drill up v hierarchii dimenzí (žádná, suma, minimum, maximum,…)

Tabulka 2-Definiční tabulka metriky

4.6.2 Členění metrik kvality dat

Metriky kvality dat můžeme podle způsobu jejich získání rozdělit do dvou skupin.

Objektivní metriky

Objektivní metriky jsou takové metriky kvality dat, které jsme schopni vždy znovu vypočítat z dat, kterých se tyto metriky týkají. Objektivní metriky jsou to často statistické charakteristiky datového souboru (průměr, rozptyl, medián, odchylky, okrajové hodnoty).

Subjektivní metriky

Naopak subjektivní metriky jsou metriky, kterými hodnotíme způsob vzniku dat, případně kvalitu jejich zdroje. Mezi subjektivní metriky patří například metriky, které hodnotí důvěryhodnost dat, dostupnost dat, stupeň utajení dat.
Proces návrhu subjektivních metrik je třeba nějakým způsobem standardizovat. To je většinou realizováno pravidly, které specifikují atributy kvality dat a postupy, které je třeba při čištění dat dodržovat. Název subjektivní metriky má svůj význam, protože tyto metriky vnikají častěji subjektivním hodnocením vlastností dat expertními uživateli, které je založeno na jejich znalostech a zkušenostech, než měřením nějakého procesu. Metriky velice často závisí na konkrétní potřebě aplikace, případně uživatele a mohou se pro různé aplikace a uživatele lišit.

4.6.3 Dimenze datové kvality

Dimenze je definována jako formalizovaná perspektiva reality a nástroj pro monitorování hodnoty metriky v daném kontextu. Dimenze datové kvality jsou definovány jako relevantní perspektivy metrik datové kvality.

Definiční tabulka dimenze

Podobně jako se definují metriky v tabulce definice metriky, všechny atributy v následující tabulce musí být definovány, aby byla dimenze správně popsána.

Dimenze	Název dimenze
ID	Jednoznačný identifikátor dimenze
Struktura dimenze	Definice hierarchie a jednotlivých levelů dimenze
Výpočty	Definice výpočtu prvků dimenze, včetně výpočtového vzorce
Zdroj	Zdroj dat pro dimenzi
Účel	Účel využití dimenze
Vlastník	Oddělení, případně člověk zodpovědný za dimenzi
Frekvence	Frekvence obnovování dat dimenze
Komentáře	Komentáře k definici dimenze

Tabulka 3-Definiční tabulka dimenze

4.6.4 Příklady dimenzí datové kvality

Mezi nejčastěji používané dimenze metrik dat dle [1] patří :

Relevantnost (Relevance)

Dimenze relevantnost popisuje, do jaké míry data splňují účel, pro který jsou používána.

Přesnost (Accuracy)

Dimenze přesnost určuje jak přesná jsou používaná data (měřeno obvykle statistickými charakteristikami pro chybu, např. směrodatná odchylka)

Včasnost (Timeliness)

Dimenze včasnost popisuje za jakou dobu lze data aktualizovat.

Dostupnost (Accessibility)

Dimenze dostupnost jak jsou již existující data dostupná. Bariéry dostupnosti mohou být technologické, např. kapacita sítě, legislativní, např. nedořešená ochrana osobních dat, či procesní, např. nevhodné či nedostatečné informace.

Porovnatelnost (Comparablity)

Dimenze porovnatelnost hodnotí možnost porovnávat, ale také spojovat data z různých zdrojů. Problémy mohou být s jednotností formátů či metod pořizování dat. Příkladem problému daného typu jsou obtíže při vytváření registru občanů (formát adresy).

Koherence (Coherence)

Dimenze koherence vyjadřuje, do jaké míry byla data vytvořena podle stejných pravidel.

Úplnost (Completeness)

Dimenze úplnost udává, jaká část potenciálních dat je zachycena v databázi, případně zda statistické charakteristiky dat nejsou ovlivněny výběrovými efekty.

Důvěryhodnost (Believability)

Dimenze důvěryhodnost určuje, jak jsou data správná a důvěryhodná.

Formát (Consistent reprezentation)

Dimenze formát popisuje, jak data odpovídají danému formátu

Snadná manipulace (Ease of manipulation)

Dimenze popisující s kterými daty se snadno manipuluje a která data lze využít za jiným účelem

Správnost (Free of error)

Dimenze popisující jak jsou data správná a hodnověrná.

Interpretace (Interpretability)
Dimenze popisující zda jsou data v daném jazyku, odpovídajících jednotkách a zda je jejich definice je jasná.

Objektivita (Objektivity)
Dimenze popisující, která data jsou nezkreslená, nepředpojatá a nezaujatá.

Reputace (Reputation)
Dimenze popisující jak data odpovídají tomu co popisují.

Bezpečnost (Security)
Dimenze popisující který přístup k datům je dostatečně bezpečný

Srozumitelnost (Understability)
Dimenze popisující jak jsou data pochopitelná.

Přidaná hodnota (Value-added)
Dimenze popisující, která data jsou prospěšná a jejichž využití přináší výhody.

4.6.5 Agregace metrik
Metriky můžeme podle stupně agregace rozdělit do dvou základních skupin. První skupinou jsou metriky detailní a druhou skupinou metriky sumární.

Detailní a sumární metriky

Detailní metriky
Detailní metriky jsou jednotlivé míry, které nám dávají základ pro kalkulaci metrik sumárních. Detailní metriky poskytují informace o specifickým problémech datové kvality a datových defektech.

Sumární metriky
Naopak sumární metriky reprezentují celkovou hodnotu (celkovou datovou kvalitu) jisté oblasti a jsou primárně založeny na business vnímání.

Příkladem sumárních metrik jsou například :

- Metriky kontroly datové kvality
- Metriky efektivity datové kvality
- Metriky vyhovění koncovým uživatelům
- Přesnost
- Kompletnost

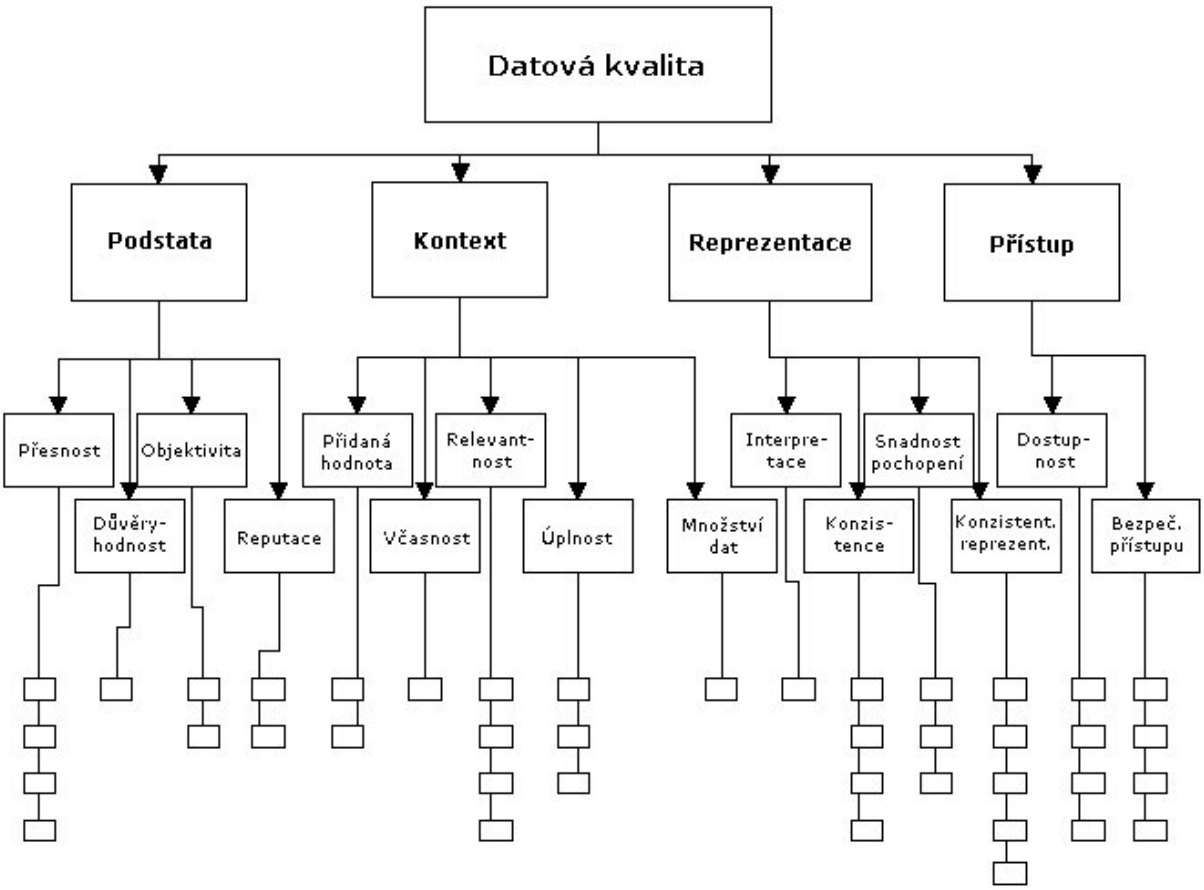
Sumární metriky jsou vypočítávány z detailních metrik relevantní oblasti jako vážený průměr a jsou normalizovány na stejnou základní čáru aby umožnily kompatibilitu v čase a oblasti.

Metriky mohou být sumarizovány na různý level, strom metrik se typicky skládá z více než jednoho levelu sumárních metrik.

Hierarchické uspořádání metrik – strom metrik

Na obrázku přeloženém z [2] je názorné zobrazení příkladu hierarchického uspořádání metrik

V spodní části jsou znázorněny hodnoty dané metriky pro jednotlivé atributy, druhý level zobrazuje dimenze těchto metrik a třetí level kategorie, kterými v tomto smyslu rozumíme jakési skupiny dimenzí.



Hierarchický strom metrik

5 Návrh základních metrik a metrik pro sledování defektů a jejich vyhodnocení

5.1 Popis problému

V předchozích kapitolách jsme se seznámili s termínem datová kvality, s programem kvality dat a podrobně jsme rozebrali první část každého takového programu dat – měření kvality dat. Protože měření datové kvality slouží především praxi, v následující části si je demonstrujeme s použitím reálných dat. Navrhneme metriky datové kvality a provedeme základní i rozšířený data profiling. Jako zdroj dat nám bude sloužit textový soubor s přehledem kontaktních údajů. Měření kvality dat budeme provádět za pomocí nástroje datové kvality SAS DataFlux.

5.2 Zdrojová data

Nejprve si přiblížíme zdrojová data nad nimiž budeme data profiling provádět.
V následující tabulce je přehled všech sloupců textového souboru kontakty.txt (tento soubor je uložen na přiloženém CD).
U každého sloupce je určen datový typ, tak jak bude daný sloupec importován do nástroje datové kvality.
U každého sloupce je také uvedeno, zda se jedná o popisný sloupec (atribut označíme A) nebo hodnotu vyjadřující nějakou hodnotu (ukazatel je označen U)

Název sloupce	**Datový typ**	**Atribut/Ukazatel**
Jméno	String	A
Příjmení	String	A
Titul	String	A
Pozice	String	A
Email	String	A
Název společnosti	String	A
Telefon	String	A
Pozice	String	A
Plat	Integer	U
Ulice	String	A
Město	String	A

Pro představu, zde uvedu také ukázku dat ze souboru kontakty.csv

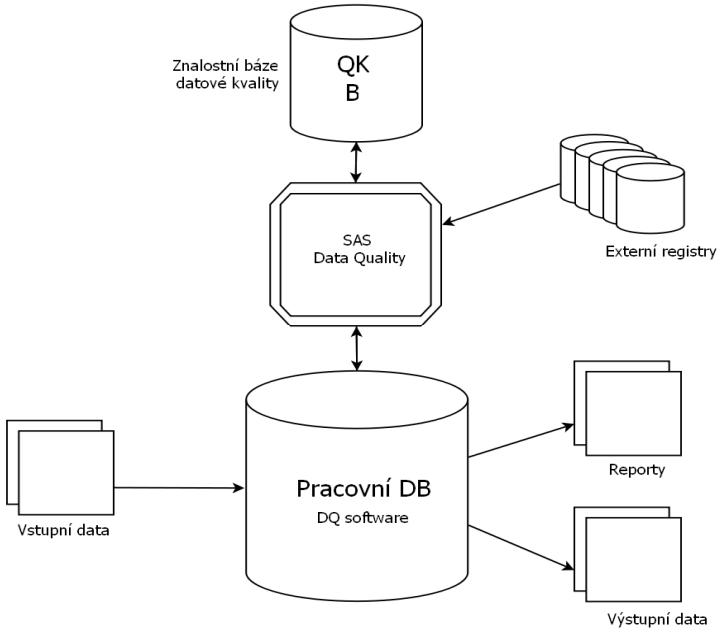
Přijmení;Jméno;Titul;Pozice;Email;Název společnosti;Telefon;Plat;Ulice;Město
Abazid;Lukáš;;projektový manager;lukas.abazid@upc.cz;UPC Česká republika, a.s.;261 107 111;36478;U Nákladového nádraží 6;PRAHA 3
Abbrent;Jan;Ing.;ředitel společnosti;abbrentj@rosshb.cz;ROSS Computers s.r.o.;569 476 400;31910;Jihlavská 893;HAVLÍČKŮV BROD
Abdul;Ivan;;ostatní;abdul@irisa.cz;IRISA, výrobní družstvo;571 484 451;22182;Jasenická 697;Vsetín
Ábel;Jan;Ing.;IT manager;abel@ans.cz;Řizení letového provozu České republiky;220373293;26113;K Letišti 1040/10;PRAHA 6

5.3 Analýza problému

Cílem je navrhnout nějaký objektivní systém pro vyhodnocování kvality dat seznamu kontaktů. Tento systém by měl být zajisté postaven na hierarchickém uspořádání metrik, které bylo podrobně popsáno v předchozí kapitole.
Pokud vezmeme v úvahu zdrojová data je jasné, že nepůjde využít všech dříve zmiňovaných dimenzí datové kvality. Jde tedy o to vybrat a nadefinovat metriky relevantní právě pro náš případ. Je zřejmé, že posuzovat např. dostupnost, bezpečnost či včasnost dat určitě nebude mít v tomto případě opodstatnění. Zaměříme se tedy na posouzení úplnosti, správnosti, správného formátu, konzistence a unifikace dat. U seznamu kontaktů se přímo nabízí také obohacení dat o pohlaví. Dále tedy můžeme vyhodnocovat úspěšnost tohohle obohacení.
Jednotlivé oblasti, které budeme posuzovat budou tedy kvalita dat, kde se zaměříme na již zmíněnou správnost, úplnost, konzistenci a formát dat. Další oblastí bude oblast možné deduplikace dat a obohacení dat.
V rámci řešení problému postupně provedeme základní profiling a rozšířený profiling dat. Analýzou jednotlivých atributů navrhneme detailní metriky a navrhneme jejich vhodnou agregaci pro celkové zhodnocení dat.

5.4 Navrhovaná architektura řešení

Jako nástroj datové kvality byl vybrán specializovaný SW pro datovou kvalitu DataFlux.dodávaný firmou SAS. My budeme využívat klientskou aplikaci dfPower Studio, sloužící především pro tvorbu business pravidel.
Návrh architektury vychází z toho, že zdrojová data jsou textovým souborem.Zdrojová data jsou tedy k DataFluxu připojena přes standardní rozhraní „Data Inputs“ jako textový soubor se záznamy oddělenými středníkem.
Jako znalostní báze obsahující ke každému sémantickému typu definice a jim odpovídající gramatiky, fonetické knihovny, knihovny regulárních výrazů, standardizační schémata a slovníky byla vybrána standardní Quality Knowledge Base verze 2009B, (dále bude označována jako QKB) dodávaná SASem včetně její české lokalizace.
Mimo této QKB budeme k ověřování zejména správnosti dat využívat také různé externí registry jako např. registr jmen a příjmení, ulic a měst.
Získané výstupy za použití nástroje kvality dat budou generovány jako textové případně html soubory a budou uloženy na přiloženém CD.



5.5 Základní profiling

Základní profiling spočívá v měření obecných charakteristik dat bez jakéhokoli business kontextu.
V rámci základního profilingu se budeme zabývat obecnými charakteristikami jednotlivých atributů, např. úplností (tj. počtem neprázdných hodnot), určením datového typu atributu, unikátností hodnot atributu atd.
Na ukazateli plat si demonstrujeme zjištění a analýzu základních číselných charakteristik.

5.6 Rozšířený profiling

V průběhu rozšířeného profilingu se budeme věnovat identifikaci datových defektů a statistikám s těmito defekty souvisejícími. Budeme se zabývat úplností dat, správností dat, jejich strukturální konzistencí a navrhneme komplexní systém metrik pro celkové posouzení datové kvality. Postupně budeme analyzovat jednotlivé atributy a pro každý navrhneme vhodné detailní metriky jeho posouzení.

5.6.1 Návrh metrik pro úplnost, správnost a formát dat - dle atributů

V seznamu je vždy uveden atribut a s ním související navrhované detailní metriky:

Atribut Jméno

1) Počet záznamů s vyplněným jménem
2) Počet správných záznamů ověřených vůči důvěryhodnému seznamu jmen

Atribut Příjmení

1) Počet záznamů s vyplněným jménem

2) Počet správných záznamů ověřených vůči důvěryhodnému seznamu příjmení

Atribut Titul

1) Počet záznamů s vyplněným titulem
2) Počet záznamů se správným titulem ověřeným vůči seznamu titulů
3) Počet záznamů, kde hodnota atributu není standardní, ale lze ji nahradit (přiřadit jí správnou hodnotu)

Atribut Pozice

1) Počet záznamů s vyplněným atributem pozice

Atribut Email

1) Počet záznamů s vyplněným e-mailem
2) Počet záznamů, kde vyplněný e-mail má správný formát

Atribut Název společnosti

1) Počet záznamů s uvedeným/neuvedeným názvem firmy
2) Počet záznamů s „odpovídajícím“ názvem firmy

Atribut Telefon

1) Počet záznamů vyplněným atributem telefon
2) Počet záznamů s telefonním číslem správného formátu

Atribut Ulice

1) Počet záznamů s vyplněným atributem Ulice

Atribut Město

1) Počet záznamů s vyplněným atributem Město
2) Počet záznamů s ověřeným atributem Město

Ukazatel plat

1) Počet záznamů s vyplněným platem
2) Počet záznamů se správným platem

5.6.2 Konzistence dat

Co se týká strukturální konzistence dat budeme posuzovat, zda je ulice konzistentní s městem, tj. určíme procento záznamů u nichž je kombinace Ulice-Město platná.

5.7 Obohacování dat

Obohacovaní dat zvyšuje informační hodnotu existujících dat přidáním nějaké dodatečné informace. V našem případě si obohacování dat demonstrujeme na příkladu určení pohlaví ze jména a příjmení.

V procesu vyhodnocení dat budeme posuzovat výsledek obohacení dat, tj. daná metrika bude udávat procento obohacených dat.

5.8 Deduplikace dat

V rámci deduplikace se budeme zabývat pouze procentuálním zjištěním počtu unifikovaných dat vůči všem datům.

5.9 Vyhodnocení kvality dat

Pro vyhodnocení kvality dat využijeme následující skupiny metrik:

- **Správnost**
- **Úplnost**
- **Konzistence**
- **Formát**

Pro vyhodnocení deduplikace dat

- **Unifikaci**

Pro vyhodnocení obohacení dat

- **Doplnění pohlaví**

Podrobný význam navržených metrik je uveden v tabulce *Navržené metriky*

Oblast	**Souhrnná metrika**	**Detailní metriky**
Kvalita dat Q1	**Správnost M1**	Udává míru zastoupení dat v databázi ověřených vůči etalonu, Je vyjádřena procentuelně:
		S1 Křestní jména ověřená vůči registru
		S2 Příjmení ověřená vůči registru
		S3 Tituly ověřené vůči seznamu titulů
		S4 Obce, které byly ověřeny vůči seznamu měst a obcí
		S5 Ověřené Názvy společnosti
		S6 Ověřené hodnoty platu
	Úplnost M2	Udává, jaká procentuální část dat jednotlivých atributů je zachycena Podrobněji viz tabulka Váhy úplností **U1-U10**
	Formát M3	Udává jaká část telefonních čísel a emailů je správného formátu Podrobněji viz tabulka Váhy formátu **F1**, **F2**
	Konzistence M4	Udává procentuální podíl konzistentních ulic a měst
Deduplikace Q2	**Unifikace**	Vyjadřuje procentuální podíl unikátních dat
Obohacení dat Q3	**Doplnění pohlaví**	Vyjadřuje procentuální podíl automaticky obohacených dat ke všem datům

Tabulka 4-Navržené metriky

5.10 Celkové vyhodnocení procesu kvality dat

Na základě výše uvedených jednotlivých detailních metrik a vah k nim přiřazených budou vypočítány dílčí a souhrnné hodnoty metrik. Součet jednotlivých vah je vždy roven 1.

Hodnota konkrétní souhrnné metriky se vypočítá podle vzorce:

$$M_x = V_1 . K_{v1} + V_2 . Kv_2 + \ldots + V_x . K_{vx}$$

kde

M_x ..... výsledná hodnota souhrnné metriky (např. hodnota metriky „správnost“)
V_x ..... procentuální výsledek konkrétního významu (např. procento validních křestních jmen)
K_{vx}..... přidělená hodnota váhy ke konkrétnímu procentuálnímu výsledku V_x v intervalu (0;1>

Výše uvedeným způsobem tedy získáme hodnoty navržených základních metrik Úplnost Správnost a Formát.

Váhy pro jednotlivé metriky budeme určovat podle důležitosti dané metriky v celkovém hodnocení procesu kvality dat. V praxi bychom tyto váhy určili za pomoci business uživatelů, kteří tato data využívají.

U seznamu kontaktů je samozřejmé, že nejdůležitější budou atributy příjmení a název společnosti, z čehož jsme schopni nejlépe daný kontakt rekonstruovat.

Tabulka *Váhy úplnosti* shrnuje jednotlivá ID detailních metrik a váhy pro výpočet souhrnné metriky Úplnost :

Úplnost	**ID metriky**	**ID váhy**	**Hodnota váhy**
Jméno	***U1***	K_{u1}	0,1
Příjmení	***U2***	K_{u2}	0,4
Titul	***U3***	K_{u3}	0
Pozice	***U4***	K_{u4}	0,05
Email	***U5***	K_{u5}	0,1
Název společnosti	***U6***	K_{u6}	0,2
Telefon	***U7***	K_{u7}	0,05
Ulice	***U8***	K_{u8}	0
Město	***U9***	K_{u9}	0,05
Plat	***U10***	K_{u10}	0,05

Tabulka 5-Váhy úplnosti

Tabulka *Váhy správnosti* shrnuje jednotlivá ID detailních metrik a jejich váhy pro souhrnnou dimenzi Správnost :

Správnost	ID metriky	ID váhy	Hodnota váhy
Jméno	***S1***	K_{s1}	0,2
Příjmení	***S2***	K_{s2}	0,4
Titul	***S3***	K_{s3}	0,05
Název společnosti	***S4***	K_{s4}	0,2
Město	***S5***	K_{s5}	0,05
Plat	***S6***	K_{u6}	0,1

Tabulka 6-Váhy správnosti

Tabulka *Váhy formátu* shrnuje jednotlivé ID metrik a jejich váhy pro dimenzi Formát :

Formát	ID metriky	ID váhy	Hodnota váhy
Email	***F1***	K_{F1}	0,5
Telefon	***F2***	K_{F2}	0,5

Tabulka 7-Váhy formátu

Nyní uvedeme definiční tabulky souhrnných metrik Úplnost, Správnost a Formát . Definiční tabulky pro jednotlivé detailní metriky, včetně konzistence, nebudu uvádět a spokojíme se pouze s jejich návrhem v rámci tabulky *Navržené metriky*.

Metrika	**Název**	**Úplnost**
	ID	M2
	Popis	Souhrnná metrika posuzující úplnost dat
Výpočet	**Jednotky**	Procenta
	Datový typ	Číslo
	Postup výpočtu	K_{u1}*U1+ K_{u2}*U2+ K_{u3}*U3+ K_{u4} *U4+ K_{u5} *U5+ K_{u6}*U6+ K_{u7}*U7+ K_{u8}*U8+ K_{u9} *U9+ K_{u10}*U10
	Rozsah hodnot	0-100%
	Podřazené metrika	Detailní metriky úplnosti pro jednotlivé atributy U1,U2,U3,U4,U5,U6,U7,U8,U9,U10
	Vztah k dimenzi	Kvalita dat
	Agregace	Suma

Tabulka 8-Definiční tabulka metriky Úplnost

Metrika	**Název**	**Správnost**
	ID	M1
	Popis	Souhrnná metrika posuzující správnost dat
Výpočet	**Jednotky**	Procenta
	Datový typ	Číslo
	Postup výpočtu	K_{s1}*S1+ K_{s2}*S2+ K_{s3}*S3+ K_{s4}*S4+ K_{s5}*S5+ K_{s6}*S6
	Rozsah hodnot	0-100%
	Podřazené metrika	Detailní metriky správnosti pro jednotlivé atributy S1,S2,S3,S4,S5,S6
	Vztah k dimenzi	Kvalita dat
	Agregace	Suma

Tabulka 9-Definiční tabulka metriky Správnost

Metrika	**Název**	**Formát**
	ID	M3
	Popis	Souhrnná metrika posuzující formát dat
Výpočet	**Jednotky**	Procenta
	Datový typ	Číslo
	Postup výpočtu	K_{f1}*F1+ K_{f2}*F2
	Rozsah hodnot	0-100%
	Podřazené metrika	Detailní metriky správnosti formátu pro jednotlivé atributy F1, F2
	Vztah k dimenzi	Kvalita dat
	Agregace	Suma

Tabulka 10-Definiční tabulka metriky Formát

Celková hodnota metriky pro jednotlivé oblasti bude pak vypočítána takto:

$$O_x = M_1 . K_{m1} + M_2 . K_{m2} + \ldots + M_x . K_{mx}$$

kde

O_x..... výsledná hodnota oblasti (např. oblasti kvality dat)
K_{mx} ..... přidělená hodnota váhy ke konkrétní oblasti v intervalu (0;1>

Jediná oblast skládající se z několika souhrnných metrik je Kvalita dat, váhy sloužící k jejímu výpočtu jsou shrnuty v tabulce *Váhy kvality dat*, stejně jako předchozí váhy jsou určeny podle business důležitosti jednotlivých souhrnných metrik.

Dílčí metrika	ID metriky	ID váhy	Hodnota váhy
Správnost	***M1***	K_{M1}	0,4
Úplnost	***M2***	K_{M2}	0,4
Konzistence	***M4***	K_{M4}	0,1
Formát	***M3***	K_{M3}	0,1

Tabulka 11-Váhy kvality dat

Definiční tabulka metriky Kvalita dat poté bude vypadat:

Metrika	**Název**	**Kvalita dat**
	ID	Q1
	Popis	Souhrnná metrika posuzující kvalitu dat
Výpočet	**Jednotky**	Procenta
	Datový typ	Číslo
	Postup výpočtu	K_{M1}*M1+ K_{M2}*M2 + K_{M3}*M3 + K_{M4}*M4
	Rozsah hodnot	0-100%
	Podřazené metrika	Souhrnné metriky Správnost, Úplnost, Konzistence, Formát M1, M2, M3, M4
	Vztah k dimenzi	Kvalita dat
	Agregace	Suma

Tabulka 12-Definiční tabulka Kvality dat

Pro přehlednost uvedeme ještě definiční tabulky jednotlivých metrik Deduplikace a Obohacení dat.

Metrika	**Název**	**Deduplikace**
	ID	Q2
	Popis	Metrika posuzující míru duplikace dat
Výpočet	**Jednotky**	Procenta
	Datový typ	Číslo
	Postup výpočtu	Metrika udává procento unifikovaných záznamů vůči všem záznamům
	Rozsah hodnot	0-100%
	Podřazené metrika	
	Vztah k dimenzi	Deduplikace
	Agregace	

Tabulka 13-Definiční tabulka Unifikace

Metrika	**Název**	**Obohacení dat**
	ID	Q3
	Popis	Metrika posuzující úspěšnost obohacení dat
Výpočet	**Jednotky**	Procenta
	Datový typ	Číslo
	Postup výpočtu	Metrika udává procento obohacených záznamů vůči všem záznamům
	Rozsah hodnot	0-100%
	Podřazené metrika	
	Vztah k dimenzi	Obohacení dat
	Agregace	

Tabulka 14-Definiční tabulka Obohacení dat

Celkový stav dat je v závěru určen hodnotami souhrnných metrik všech tří oblastí - kvalitou dat, mírou deduplikace a úspěšností obohacení dat.

Dílčí metriky *Kvality dat* jsou na sobě samozřejmě funkčně závislé, metriky správnost, konzistence i formát samozřejmě závisí na metrice úplnost.
Metrika *Unifikace* je závislá na hodnotách metrik kvality dat (čím vyšší je kvalita dat, tím vyšší je podíl automaticky sloučených a unikátních dat).

6 Realizace navrženého systému vyhodnocení kvality dat

Cílem realizace navrženého systému je výpočet jednotlivých navržených detailních metrik, souhrnných metrik a oblastí a celkové vyhodnocení kvality dat zkoumaného vzorku včetně analýzy nejslabších článků.

6.1 Základní profiling

Nejprve provedeme základní profiling, jehož cílem bude zjistit základní charakteristiky dat – zejména úplnost a unikátnost a demonstrujeme na ukazateli plat možnosti profilingu pro číselná data. K tomuto základnímu profilingu využijeme část nástroje datové kvality nazývané Profile Configurator a Profile Viewer, výsledné reporty budeme generovat do excelovských souborů.

6.1.1 Základní charakteristiky dat

V rámci základního profilingu nás zajímá zejména úplnost každého z atributů, nicméně určíme i ostatní základní standardní charakteristiky atributů jako jsou například unikátní hodnoty, datové typy, délky datového typu string apod..

Úplnost dat

Tabulka *Profiling úplnosti dat* zobrazuje pro každý atribut počet záznamů, počet null hodnot pro ukazatel plat (datový typ integer), počet prázdných hodnot pro textové atributy.

Název pole	Počet hodnot	Počet NULL hodnot	Počet prázdných hodnot
Email	16275	0	1985
Jméno	16275	0	451
Město	16275	0	164
Název společnosti	16275	0	114
Plat	16275	0	(nelze provést)
Příjmení	16275	0	5
Ulice	16275	0	195
Pozice	16275	0	5538
Telefon	16275	0	2483
Titul	16275	0	11644

Tabulka 15-Profiling úplnosti dat

Unikátní hodnoty

V tabulce *Unikátnost dat* jsou potom shrnuty počty jednoznačných záznamů každého sloupce, včetně jejich procentuelního vyjádření.

Název pole	Počet unikátních hodnot	Unikátnost
Email	11411	70,11
Jméno	946	5,81
Město	1262	7,75
Název společnosti	4491	27,59
Plat	12567	77,24
Příjmení	8220	50,51
Ulice	4567	28,06
Pozice	209	1,28
Telefon	7596	46,67
Titul	47	0,29

Tabulka 16-Unikátnost dat

Podrobnější profiling

Tabulka *Podrobnější profiling* ukazuje další zajímavé charakteristiky jako je např. analýza primárních klíčů. Zde vidíme, že žádný ze sloupců není vhodným kandidátem. Také je provedena analýza datového typu a délky pro datový typ STRING. Poslední údaj se týká analýzy pozice sloupce ve zdroji.

Název pole	Kandidát na primární klíč	Datový typ	Délka	Pozice
Email	no	STRING	35 chars	5
Jméno	no	STRING	14 chars	2
Město	no	STRING	27 chars	10
Název společnosti	no	STRING	44 chars	6
Plat	no	INTEGER	0 chars	8
Příjmení	no	STRING	16 chars	1
Ulice	no	STRING	37 chars	9
Pozice	no	STRING	28 chars	4
Telefon	no	STRING	15 chars	7
Titul	no	STRING	5 chars	3

Tabulka 17-Podrobnější profiling

6.1.2 Základní profiling platu

Ukazatel Plat je jediným číselným ukazatelem v našem seznamu kontaktů, proto si právě na něm demonstrujeme základní profiling číselné hodnoty.
Je ovšem třeba vzít v úvahu, že ukazatel plat byl vygenerován generátorem náhodných čísel s normálním rozdělením a pouze částečně ručně a automaticky upraven, proto také statistické rozdělení hodnot ukazatele se bude významně lišit od rozdělení reálného, což bude patrné zejména na směrodatné odchylce, střední chybě, frekvenční analýze a percentilech.

Základní profiling sloupce

Součástí tzv. základního profilingu platu je zjištění základních statistik. Tyto základní statistiky shrnuje tabulka *Základní profiling platu*.
Originální výstup profilingu sloupce vyexportovaný z DataFluxu do souboru xls je uložen v souboru Základní profiling platu.xls.

Název pole	Plat
Pozice	10
Počet	16275
Počet NULL hodnot	0
Počet prázdných hodnot	(not applicable)
Minimální hodnota	0
Maximální hodnota	999999999
Mode	40000
Pattern Count	(not applicable)
Unique Count	56
Uniqueness	0,34
Kandidát na primární klíč	no
Datový typ	INTEGER
Délka	0 chars
Aktuální typ	integer
Minimální délka	(not applicable)
Maximální délka	(not applicable)
Průměr	540144,3006
Medián	42000
Počet not NULL hodnot	16275
Nullable	UNKNOWN
Desetinná místa	0
Směrodatná odchylka	22178960,3
Střední chyba	173852,3978
Procento Null hodnot	0

Tabulka 18-Zákaldní profiling platu

Percentily

Analýza percentilů s intervalem percentilu =10 porovnávajících platy mezi sebou 10 je uložena v souboru Percentily.xls.

Interval	Hodnota
10	28000
20	32000
30	35000
40	39000
50	42000
60	46000
70	50000
80	53000
90	57000

Tabulka 19-Percentily platu

Analýza krajních hodnot

Analýza krajních hodnot slouží k zjištění minimálních a maximálních daného ukazatele. Výsledek analýzy je uložen v souboru Krajní hodnoty platu.xls.

Minimální hodnoty	Maximální hodnoty
0	100000
1	110000
15000	170000
16000	99999999
17000	999999999

Tabulka 20-Krajní hodnoty platu

Frekvenční analýza

Následující tabulka shrnuje výsledek frekvenční analýzy, z výsledků analýzy je patrné, že hodnota atributu plat byla generována jako náhodné číslo s normálním rozdělením a byly provedeny jen mírné ruční úpravy.
Výstup analýzy je uložen v souboru Frekvenční analýza platu.xls.

Hodnota	Počet	Procenta
40000	499	3,07
58000	492	3,02
28000	481	2,96
30000	476	2,92
55000	472	2,9
36000	471	2,89
31000	470	2,89
43000	469	2,88
50000	465	2,86
52000	463	2,84
38000	463	2,84
49000	462	2,84
39000	462	2,84
56000	455	2,8
32000	454	2,79
45000	454	2,79
54000	453	2,78
47000	452	2,78
35000	449	2,76
53000	449	2,76

Tabulka 21-Frekvenční analýza platu

6.2 Kvalita dat

6.2.1 Ukazatel plat

Počet záznamů s "předpokládaným správným" platem, jako omezující podmínky omezíme plat zdola částkou minimální mzdy tj. 8000Kč a shora částkou 300 000Kč
Výpočet provedeme za pomoci skriptu plat.sql za pomoci databáze Oracle.

Tabulka *Vyhodnocení platu* obsahuje výsledné hodnoty metrik souvisejících s daným sloupcem.

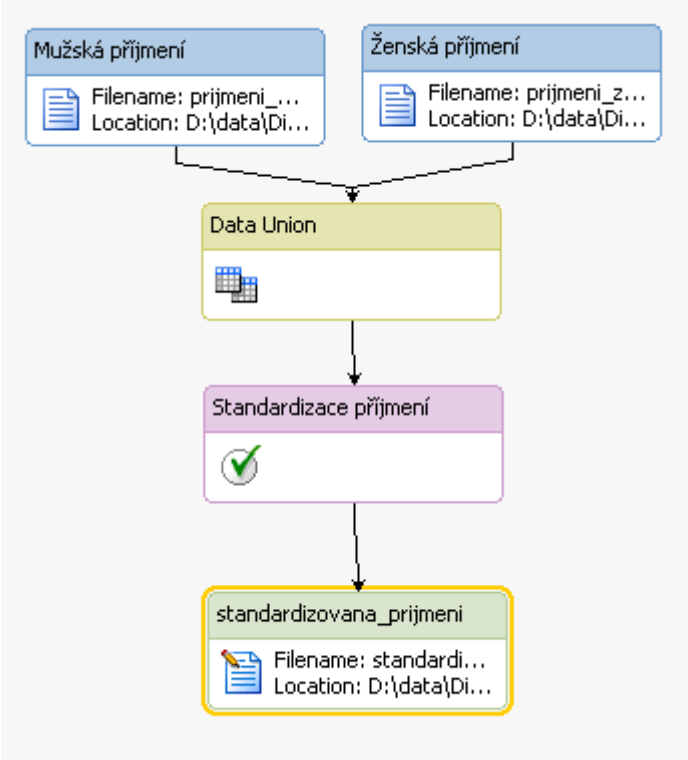
	Metrika	**Hodnota**
Plat	Počet hodnot	16275
	Počet vyplněných hodnot	16275
	Počet prázdných hodnot	**0**
	Počet správně vyplněných hodnot	**16259**

Tabulka 22-Vyhodnocení platu

6.2.2 Atribut příjmení

Pro posouzení správnosti příjmení se nabízí jednoduché řešení, využít standardního nástroje SW datové kvality (ať už se jedná o nástroj Gender analysis, případně Identifikation analysis). Bohužel využití ani jednoho z těchto nástrojů za využitích různých definic nepřináší v českém prostředí uspokojivé výsledky, stále existují správná příjmení která nebyla verifikována. Proto nezbývá než pátrat po jiném řešení a tím je porovnávat příjmení přímo vůči nějakému důvěryhodnému registru.
Jako vhodný registr se ukázal zdroj s přehledem frekvencí všech příjmení dodávaných Ministerstvem vnitra [6] a to speciálně s příjmeními všech osob s platným pobytem v ČR. Do této databáze jsou vybírány pouze osoby živé s platným pobytem v ČR a to ke dni 15.07.2009, výběr obsahuje všechna příjmení včetně chybně zavedených příjmení.
Prvním krokem bude vytvoření databáze standardizovaných příjmení, protože ani data Ministerstva vnitra se neukázala jako stoprocentně kvalitní provedeme v rámci přípravy nejenom sloučení souborů s ženskými a mužskými příjmeními ale také standardizaci, tj. očištění od mezer a diakritiky. Výstupem těchto akcí je soubor standardizovana_prijmeni.csv.

Schéma *Vytvoření registru příjmení* znázorňuje řešení pomocí DataFluxu:



Scjéma vytvoření registru příjmení

Protože tento soubor je plný duplicit provedeme na něj select distinct, jehož výstupem bude soubor pomocí něhož budeme správnost dat ověřovat (standardizovana_prijmeni_distinct.txt).

Při ověřování provedeme nejprve standardizaci příjmení ze seznamu kontaktů (využijeme stejné standardizační definice) a poté pomocí left joinu ke každému příjmení doplníme jeho standardizovaný tvar z registru, pokud existuje.

Doplníme proměnnou Ověřeno pro posouzení správnosti dat a uděláme její frekvenční analýzu.
Schéma ověření příjmení znázorňuje řešení ověření jako job v DataFluxu:

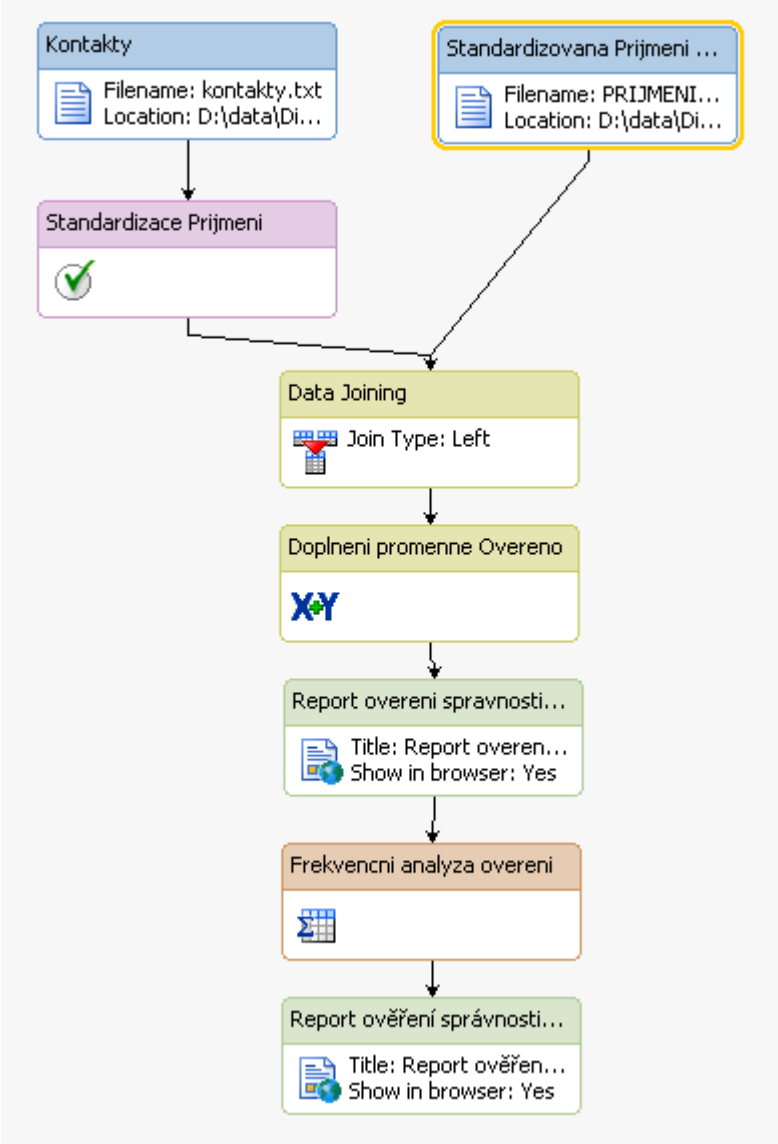


Schéma ověření příjmení

Výstupem ověřování je html soubor Přehled ověření správnosti příjmení.html zobrazující ke každému příjmení jeho standardní tvar, standardní tvar z registru a zda se hodnotu podařilo ověřit.

Report ověření správnosti příjmení

Ověřeno	Počet
false	1173
true	15102

Výsledky měření správnosti a úplnosti příjmení jsou v tabulce *Vyhodnocení příjmení* :

	Metrika	Hodnota
Příjmení	Počet hodnot	16275
	Počet vyplněných hodnot	16270
	Počet prázdných hodnot	**5**
	Počet správně vyplněných hodnot	**15102**

Tabulka 23-Vyhodnocení příjmení

6.2.3 Atribut jméno

Podobně jako v případě příjmení budeme i jméno ověřovat vůči seznamu jmen získaných ze stránek Ministerstva vnitra [6], nástroje standardně dodávané se i v tomto případě ukázaly jako nevyhovující.

Opět nejprve obdobným způsobem vytvoříme soubor se seznamem jmen se kterým budeme data porovnávat. Název souboru pouze se standardizovanými daty je standardizovana_jmena.txt, soubor s již deduplikovanými daty vhodnými k ověření potom standardizovana_jmena_distinct.txt

Výstupem ověřování je html soubor Přehled ověření správnosti jmen.html zobrazující ke každému jménu jeho standardní tvar, standardní tvar z registru a zda se hodnotu podařilo ověřit.

Přehled ověření správnosti jmen.html

Jméno	Jméno Standard	Jméno Standard Z Registru	Ověřeno
			false
Sybase Francie		SYBASE FRANCIE	false
Posedníková		POSEDNIKOVA	false
???			false
Jerden Van		JERDEN VAN	false
Lukáš	LUKAS	LUKAS	true
Jan	JAN	JAN	true
Ivan	IVAN	IVAN	true
Jan	JAN	JAN	true
Titus	TITUS	TITUS	true
Martin	MARTIN	MARTIN	true
Jan	JAN	JAN	true
Miloš	MILOS	MILOS	true
Tomáš	TOMAS	TOMAS	true
Zbyněk	ZBYNEK	ZBYNEK	true

Report ověření správnosti jmen.html

Ověřeno	Počet
false	702
true	15573

Výsledky měření správnosti a úplnosti jména jsou uvedeny v tabulce *Vyhodnocení jména*.

	Metrika	Hodnota
Jméno	Počet hodnot	16275
	Počet vyplněných hodnot	15824
	Počet prázdných hodnot	**451**
	Počet správně vyplněných hodnot	**15573**

Tabulka 24-Vyhodnocení jména

6.2.4 Atribut titul

Pro ověření správnosti titulu se nabízí využít standardního nástroje Data validation. Nicméně nikde v QKB se nepodařil najít vhodný registr k porovnání titulů, proto byl vytvořen takový registr za pomocí internetové encyklopedie Wikipedia [4].
Zdrojový textový soubor pomocí něhož budeme ověřovat je akademické_tituly.txt.

Počet záznamů se správným titulem ověřeným vůči seznamu titulů

Počet správných hodnot atributu titul ověříme vůči vytvořenému seznamu titulů.V následující tabulce vidíme distribuci ověřených titulů, hodnota Nepřiřazeno značí, že se titul nepodařilo ověřit vůči zdroji.

Titul	Počet hodnot
Bc.	34
CSc.	2
Csc.	1
Doc.	1
Dr.	53
Ing.	3830
JUDr.	33
MBA	4
MUDr.	10
MVDr.	8
Mgr.	209
PHDr.	9
PhDr.	4
Prof.	7
RNDr.	148
Nepřiřazeno	**11922**

Tabulka 25-Ověření titulů

Nyní podrobněji rozebereme základní statistiky atributu titul :

	Metrika	Hodnota
Titul	Počet hodnot	16275
	Počet vyplněných hodnot	4631
	Počet prázdných hodnot	**11644**
	Počet správně vyplněných hodnot	**4353**

Tabulka 26-Titul bez standardizace

Je zřejmé, že se nám podařilo ověřit většinu vyplněných hodnot už bez jakékoli standardizace.

Počet záznamů, kde hodnota atributu není standardní, ale lze ji nahradit (přiřadit jí správnou hodnotu)

Ze základního profilingu atributu titul vidíme, že počet unikátních hodnot atributu je poměrně nízký (47). Provedeme tedy frekvenční analýzu hodnot atributu. Report z této analýzy je uložen v souboru Titul Frequency Distribution.html.
Z analýzy je zřejmé, že mimo vyloženě nepřípustných hodnot atributu existují i hodnoty, pro něž lze standardní tvar nalézt, proto navrhneme pravidla, pomocí nichž toho docílíme. Tyto pravidla jsou uloženy v souboru titul_schema.txt.

Z procesu standardizace získáme hodnotu metriky počet záznamů, kde hodnota není standardní ale lze ji nahradit (odpovídají záznamům s hodnotou TRUE v souboru Titul Standardization Report.html) :

Počet nahrazených hodnot	**28**

Na standardizovaných titulech provedeme ověření vůči seznamu titulů. Tabulka zobrazuje frekvenční analýzu, kde hodnota Nepřiřazen značí, že se nenašel vzor (počet hodnot Nepřiřazen je tak vysoký, protože obsahuje i prázdné hodnoty) :

Standardní tvar titulu	Počet hodnot
Bc.	34
CSc.	2
Csc.	1
Doc.	7
Dr.	54
Ing.	3846
JUDr.	33
MBA	4
MUDr.	10
MVDr.	8
Mgr.	213
PHDr.	9
PhDr.	4
Prof.	7
RNDr.	148
Nepřiřazen	11895

Tabulka 27-Standardní tvar titulu

Následující tabulka shrnuje výsledný stav po provedení standardizace titulů:

	Metrika	Hodnota
Titul	Počet hodnot	16275
	Počet vyplněných hodnot	4631
	Počet prázdných hodnot	**11644**
	Počet správně vyplněných hodnot	**4380**

Tabulka 28-Vyhodnocení titulu

6.2.5 Atribut pozice

Ověřením správnosti atributu pozice se nebudeme zabývat, bylo by třeba získat nějaký seznam pozic v zaměstnání a vůči němu ověřovat, ale pro seznam kontaktů není tahle informace příliš zajímavá.

Pro atribut pozice provedeme tedy pouze výpočet metriky úplnost, který shrnuje tabulka *Vyhodnocení pozice*

	Metrika	Hodnota
Pozice	Počet hodnot	16275
	Počet prázdných hodnot	**5538**

Tabulka 29-Vyhodnocení pozice

6.2.6 Atribut email

Ověřování správnosti formátu atributu Email budeme provádět na základě validace formátu emailu. Provedeme parsování emailu do jednotlivých tokenů, pokud se nám email podaří v pořádku rozparsovat budeme jej považovat za správný, v opačném případě označíme formát za nesprávný.
Parsování emailu provedeme za pomocí standardního nástroje parsingu v nástroji kvality dat. Analýzou a vyhodnocením různých parsovacích definic jsem jako nejlepší vyhodnotila standardní definici E-Mail z QKB.
Postup řešení je tedy následující, nejprve načteme zdrojový soubor, poté provedeme parsování emailu dle již zmíněné definice a parsování pomocí standardních Data Outputs vyhodnotíme.

Následující diagram popisuje proces validace emailů pomocí jobů v DataFluxu:

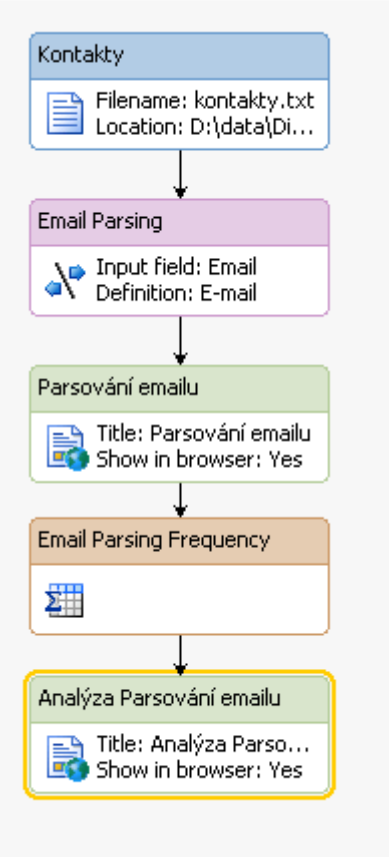


Schéma parsingu emailu

Výsledek parsingu emailů je uložen v souboru Parsování emailu.html. Kde pro každý email vidíme jeho rozdělení do jednotlivých tokenů - Mailbox, Sub-Doména a Top-Level Domain. Ve sloupci Výsledek je zobrazen výsledek parsingu. Hodnota OK odpovídá úspěšnému ověření tvaru emailu, hodnota NO SOLUTION neúspěšnému ověření případně prázdné hodnotě atributu.
Ukázka z Parsování emailu.htm :

alexej.vitek	setuza	cz	OK
alexej.vitek	setuza	cz	OK
alabaar	hotmail	com	OK
libuse.al-sharibatyova	czech-tv	cz	OK
altmanm	logica	com	OK
alturbanova	kvk	cz	OK
ambersky	direct21	sk	OK
martin.ambler	ca	com	OK
P_Ambros	lineanivnice	cz	OK
marketing	lineanivnice	cz	OK
ambroz.antonin	cpost	cz	OK
			NO SOLUTION
ambroz.pf	allianz	cz	OK
ambroz.pf	allianz	cz	OK
ambroz	datasys	cz	OK
andel	ucl.cas	cz	OK
josef.andel	net4net	cz	OK

Frekvenční analýzou atributu Výsledek získáváme tabulku *Vyhodnocení parsingu emailu*.

Výsledek	Počet
NO SOLUTION	2330
OK	13941

Tabulka 30-Vyhodnocení parsingu emailu

Tabulka *Vyhodnocení emailu* shrnuje výsledek profilingu atributu Email :

	Metrika	Hodnota
Email	Počet hodnot	16275
	Počet vyplněných hodnot	14290
	Počet prázdných hodnot	**1985**
	Počet hodnot správného formátu	**13941**

Tabulka 31-Vyhodnocení emailu

6.2.7 Atribut název společnosti

K ověření správnosti záznamů v poli Název společnosti využijeme tzv. identifikační analýzy, která na základě různých pravidel rozpoznává, zda obsah pole odpovídá firmě, jednotlivci případně hodnotu nelze rozpoznat. Tato analýza sice není nejspolehlivější, nicméně pro naše potřeby se jeví jako dostatečná.

Výstupem identifikační analýzy je soubor Ověření názvu společnosti.htm.

Ověření názvu společnosti

Název Společnosti	Ověření Názvu Společnosti
Sybase Products Central Europe	UNKNOWN
Toyota Peugeot Citroen Automobile Czech,	ORGANIZATION
Demag Delaval Industrial Turbomachinery	UNKNOWN
LIHOVAR CHLUM s.r.o.	ORGANIZATION
	UNKNOWN
UPC Česká republika, a.s.	ORGANIZATION
ROSS Computers s.r.o.	ORGANIZATION
IRISA, výrobní družstvo	ORGANIZATION
Řízení letového provozu České republiky	ORGANIZATION
ABB LUMMUS GLOBAL s.r.o.	ORGANIZATION

Ve sloupci Název společnosti je původní název ze zdrojového souboru ve sloupci Ověření názvu společnosti potom výsledek ověření.

Tabulka *Ověření názvu společnosti* shrnuje a vysvětluje hodnoty tohoto sloupce.

Individual	Jednotlivec
Organization	Organizace
Unknown	Nepodařilo se ověřit

Tabulka 32-Identifikace názvu společnosti

Výsledek frekvenční analýzy hodnot je uložen v souboru Analýza ověření názvu společnosti.htm.

Analýza ověření názvu společnosti

Ověření Názvu Společnosti	Počet
INDIVIDUAL	567
ORGANIZATION	15211
UNKNOWN	497

Z výsledku analýzy jasně vidíme, že většina záznamů se podařila jako firma ověřit. Pro naše potřeby budeme za správně ověřené hodnoty považovat pouze hodnoty ORGANIZATION.

Následující tabulka shrnuje profiling atributu Název společnosti :

	Metrika	**Hodnota**
Název společnosti	Počet hodnot	16275
	Počet vyplněných hodnot	16161
	Počet prázdných hodnot	**114**
	Počet správně vyplněných hodnot	**15211**

Tabulka 33-Vyhodnocení názvu společnosti

6.2.8 Atribut telefon

Pro získání počtu čísel správného formátu nejeví jako nejvýhodnější využít standardního parsingu telefonního čísla pomocí nástroje Parsing s definicí Phone v české lokalizaci, tato definice kontroluje totiž také například délku telefonního čísla.
Report obsahující výsledek parsingu telefonních čísel je Parsování telefonního čísla.htm.
V tomto reportu je zobrazeno původní telefonní číslo a jeho rozdělení do jednotlivých tokenů. V sloupci telefon je původní neupravované číslo ze zdroje, dále report zobrazuje tokeny obsahující předvolbu země, základ telefonního čísla a klapku.
Pro nás rozhodující je sloupec Výsledek, který nám říká, zda se dané telefonní číslo povedlo rozparsovat do jednotlivých tokenů a jeho formát je tedy správný. Hodnoty atributu Výsledek jsou OK – pokud se povedlo najít odpovídající tvar a NO SOLUTION – pokud formát čísla není správný.

Parsování telefonního čísla

Telefon	Předvolba Země	Základní Telefonní Číslo	Klapka	Výsledek
				NO SOLUTION
261 107 111		261 107 111		OK
569 476 400		569 476 400		OK
571 484 451		571 484 451		OK
220373293		220373293		OK
05/45517488		545517488		OK
+421 2 59984 22	+42	1 2 59984 22		OK
326 819 228		326 819 228		OK
				NO SOLUTION
				NO SOLUTION
541 637 389		541 637 389		OK
274 813 781		274 813 781		OK

Frekvenční analýza sloupce Výsledek je uložena v reportu Analýza parsování telefonního čísla.htm, který obsahuje pro každou hodnotu atributu její počet.

Analýza parsování telefonního čísla

Výsledek Parsování	Počet
NO SOLUTION	4357
OK	11914

Pro celkový počet ověřených formátů čísel musíme hodnoty přepočítat, protože parsovací algoritmus započítává mezi neúspěšné i nevyplněné hodnoty.

	Metrika	Hodnota
Telefon	Počet hodnot	16275
	Počet vyplněných hodnot	13792
	Počet prázdných hodnot	**2483**
	Počet správně vyplněných hodnot	**11914**

Tabulka 34-Vyhodnoceni telefonu

6.2.9 Atribut město

Po analýze se opět ukázalo jako nejlepší nevyužívat standardních funkcionalit nástroje datové kvality, ale využít porovnání s nějakým důvěryhodným zdrojem.
Jako vyhovující zdroj se ukázal registr měst vytvořený ze seznamu PSČ měst a obcí, který byl získán z internetového zdroje [7], vůči němu budeme také následně správnost atributu město ověřovat.
Upravený soubor vhodný pro import do DataFluxu je uložen pod názvem obce.txt.
Po první validaci dat (report Validace města.htm) je zřejmé, že nevalidovaných dat je velké množství, proto bude nutné do validačního souboru přidat zejména části Prahy a také provést parsing záznamů město s pomlčkou a brát v úvahu pouze token před pomlčkou, který zpravidla obsahuje obec.

Výstup první validace města je uložen v souboru Validace města.htm.

Validace města

Město	Hodnota
	FAIL
PRAHA 3	FAIL
HAVLÍČKŮV BROD	
Vsetín	
PRAHA 6	FAIL
Brno	
BRATISLAVA	FAIL
Mladá Boleslav	
Olomouc	
Praha 1	FAIL
BRNO	
Praha 10	FAIL
Praha 10	FAIL
PRAHA 4	FAIL
PRAHA 4	FAIL

Frekvenční analýza poté v souboru Analýza Validace města.htm

Analýza Validace města

Hodnota	Počet
FAIL	10829

Po přidání částí Prahy do validačního souboru získáme výstupní report Validace města po přidání částí Prahy.htm a frekvenční analýzu Analýza validace města po přidání částí Prahy.htm.

Analýza validace města po přidání částí Prahy

Hodnota	Počet
FAIL	2714

Dále vytvoříme parsovací pravidlo Město_Parsing, které nám rozdělí město do tokenů Město a část. Výstup parsingu je uložen v souboru Mesto_Parsing.txt

Po rozparsování Atributu Město dle pomlčky dostáváme výsledný report Validace města po parsování.htm a frekvenční analýzu Analýza validace města po parsování.htm.

Analýza validace města po parsování

Město	Ověření
FAIL	1845

Následující tabulka celkově shrnuje profiling atributu Město :

	Metrika	Hodnota
Město	Počet hodnot	16275
	Počet vyplněných hodnot	16111
	Počet prázdných hodnot	**164**
	Počet správně vyplněných hodnot	**14430**

Tabulka 35-Vyhodnocení města

6.2.10 Konzistence dat

Analýzou problému se podařilo zjistit, že používaný nástroj datové kvality sice obsahuje funkcionalitu pro verifikaci daného adresního místa, bohužel tahle verifikace je dodávána externím dodavatelem a není pro české prostředí standardně licencována a není ji tedy možné využít. Proto je třeba nalézt jiné řešení jak posoudit konzistenci měst a ulic, tj. zda zadaná kombinace skutečně existuje a odpoví dá si.
Pátrání po internetových zdrojích bohužel skončilo u verifikace obcí, databáze všech českých ulic se mi nepodařilo sehnat.
Databázi všech českých adres spravuje a poskytuje Ministerstvo vnitra jako Centrální databázi Územně identifikačního registru. Tuhle databázi třetím stranám zpřístupňuje pomocí aplikace UIR-ADR, obsahující potřebné informace např. také ve formě CSV souborů. Těchto CSV souborů využijeme a za pomocí nich si vytvoříme registr obcí a jim odpovídajících ulic. Tento registr je uložen pod názvem obec_ulice.txt.
Postup řešení jsem zvolila následující. Protože předpokládáme různé nestandardní zápisy měst a ulic provedeme před porovnáním ještě parsing a standardizaci seznamu ulic a měst ze

souboru kontakty.txt a stejně tak provedeme standardizaci ulic a obcí z registru obec_ulice.txt a až teprve dojde k samotnému ověřování dat, bez těchto kroků by se nám bohužel velice často nepodařilo ověřit ani správnou kombinaci

Parsing měst a ulic

Nejprve provedeme parsing sloupce Město a to tak, že pomocí mírně upravené parsovací definice City-State/Province-Postal z QKB oddělíme do sloupce df_obec název v obce ze sloupce Město.
Obdobně pomocí mírně upravené parsovací definice Adress (2009A) z QKB oddělíme do pole df_ulice název ulice.

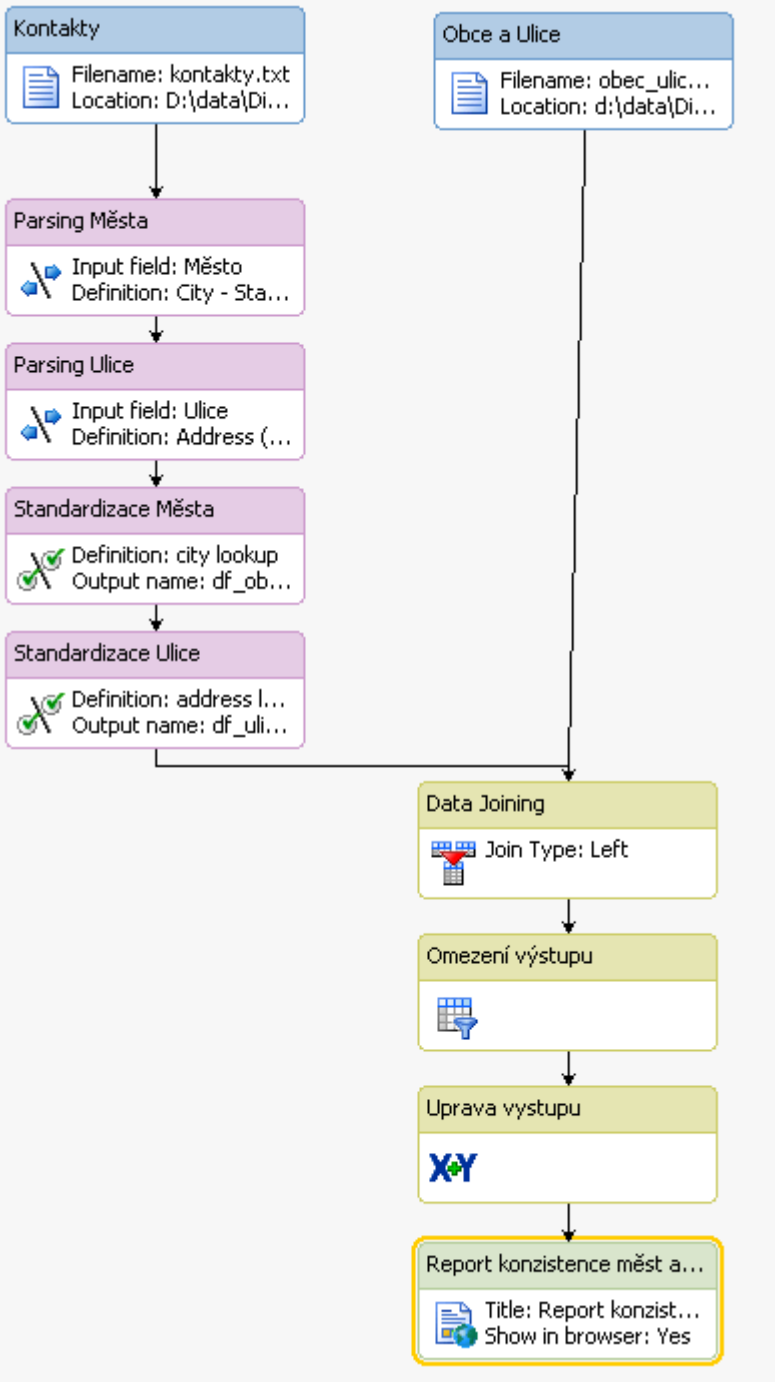
Standardizace měst a ulic

Pomocí nástroje standardizace za využití nově vytvořených schémat city_lookup a adress_lookup, která obsahují jednotlivé standardní tvary pro konkrétní česká města a ulice, dohledáme pro dané město resp. ulici standardní tvar. Výhodou využití těchto modifikovaných standardních schémat je to, že nalezneme vzor v registru i u ne úplně správně zadaných dat. Tedy např. k ulici „Masarykovo nám.“ přiřadíme „Masarykovo náměstí“ atd.

Porovnání s registrem

Spojením tabulek Kontaktů a Registru obcí a ulic left joinem přes standardní tvar města a ulice získáme výstupní tabulku obsahující pro každý standardní tvar města a ulic dohledaný tvar v Registru, pokud takový tvar existuje. V opačném případě hodnotu NULL. Poté doplníme výstup o proměnnou Ověřeno, která bude obsahovat hodnotu true pokud se data podařilo vůči registru ověřit a false pokud tomu tak nebylo.

Následující schéma znázorňuje řešení pomocí Jobu v DataFluxu:



Shéma kontroly konzistence dat

Výstupy

Reporty

Výstupem měření konzistence je html Report konzistence měst a ulic.html, který pro každé Město a Ulici obsahuje informace dohledané v registru. Z reportu je patrné, že tohle dohledání funguje i pro ne zcela správné tvary ulic a měst. Jako např. pro vzor „Václava Klementa“ byl správně nalezen a ověřen vzor „tř. Václava Klimenta“ . Dále se dá z reportu vysledovat i to, že přiřazení se velmi často nepodaří proto, že není správně vyplněna ulice, zejména u menších

obcí, kde by v ulici měla být zadána část obce a v obci samotné sídlo obce, ale častokrát se to tak neděje.
Ukázka z reportu Report konzistence měst a ulic.html:

HAVLÍČKŮV BROD	Jihlavská 893	Havlíčkův Brod	Jihlavská	true
Vsetín	Jasenická 697	Vsetín	Jasenická	true
PRAHA 6	K Letišti 1040/10	Praha	K letišti	true
Brno	Milady Horákové 13	Brno	Milady Horákové	true
BRATISLAVA	Mostová 2			false
Mladá Boleslav	Václava Klementa 869	Mladá Boleslav	tř. Václava Klementa	true
Olomouc	Březinova 7	Olomouc	Březinova	true
Praha 1	Václavské náměstí 56	Praha	Václavské náměstí	true
BRNO	Dřevařská 11	Brno	Dřevařská	true
Praha 10	Průběžná 85	Praha	Průběžná	true
Praha 10	Průběžná 85	Praha	Průběžná	true
PRAHA 4	Poláčkova 1976/2	Praha	Poláčkova	true
PRAHA 4	Želetavská 1448/7	Praha	Želetavská	true
BRATISLAVA	Mostová 2			false
LETOHRAD	Šedivská 339	Letohrad	Šedivská	true
BLANSKO	Poříčí 24	Blansko	Poříčí	true
PRAHA 4	Donovalská 808/17	Praha	Donovalská	true

Z reportu je patrné, že k většině měst a ulic se standardní tvary podařily dohledat, výjimkou je např. záznam s městem Bratislava, ten se nedohledal, protože porovnáváme pouze vůči českému registru.

Shrnující tabulka

Frekvenční analýzou sloupce Ověřeno do reportu Report konzistence měst a ulic.html získáváme :

Frekvenční analýza konzistence Měst a Ulic

Overeno	Overeno_count
false	3173
true	13160

	Metrika	Hodnota
Konzistence	Počet hodnot	16275
	Počet konzistentních hodnot	13160

Tabulka 36-Vyhodnocení konzistence

6.3 Deduplikace

Metrika počet unifikovaných záznamů nás bude zajímat jen do té míry, abychom zjistili počet takovýchto záznamů. Nebudeme se nijak věnovat řešením duplicit dat, hledáním nejlepšího master záznamu apod.
Pomocí standardního nástroje DataFluxu na základě podobnosti jmen a příjmení určíme počet kontaktů, které se vyskytují v seznamu vícekrát a počet kontaktů, které se vyskytují pouze jednou.

Výstupem téhle analýzy jsou soubory duplikovane_zaznamy.txt a neduplikovane zaznamy.txt.
V tabulce *Duplicity* jsou uvedeny počty skupin záznamů které se opakují a počet záznamů, které se vyskytují právě jednou.

Počet záznamů, které se opakují	Počet neduplicitních záznamů
2214	10252

Tabulka 37-Duplicity

	Metrika	Hodnota
Deduplikace dat	Počet hodnot	16275
	Počet unifikovaných hodnot	12466

Tabulka 38-Vyhodnocení deduplikace

6.4 Obohacení dat

Jako obohacení dat provedeme tzv. gender analýzu a určíme ke každému záznamu pohlaví. K určení využijeme sloupce jméno a přímení a jejich porovnání s QKB.

Doplnění pohlaví

Doplnění pohlaví se bude skládat ze dvou dílčích kroků. V prvním kroku pomocí SAS DataFluxu určíme dílčí údaje, tj. zjistíme pohlaví odpovídající pouze jménu, případně příjmení. V druhém kroku za pomoci databáze uděláme merge těchto hodnot dle zadaných pravidel.

Určení dílčích hodnot

Ze sloupce Jmeno získáme výstup Jmeno_gender s hodnotami ‘U‘ (nezjištěno), ‘F‘ (žena),‘M‘ (muž) .
Ze sloupce Prijmeni získáme výstup Prijmeni_gender ‘U‘, ‘F‘,‘M‘ .

Protože předpokládáme, že jednotlivá pohlaví získaná ze jména a příjmení se mohou v některých případech lišit, výstupem této první části doplnění pohlaví bude textový soubor kontakty_gender.txt se sloupci Jmeno, Prijmeni, Jmeno_gender, Prijmeni_gender, který podrobíme následující analýze.

Jmeno_gender	Prijmeni_gender	Pohlaví
F	F	F
F	U	F
U	F	F
M	M	M
U	M	M
M	U	M
U	U	U

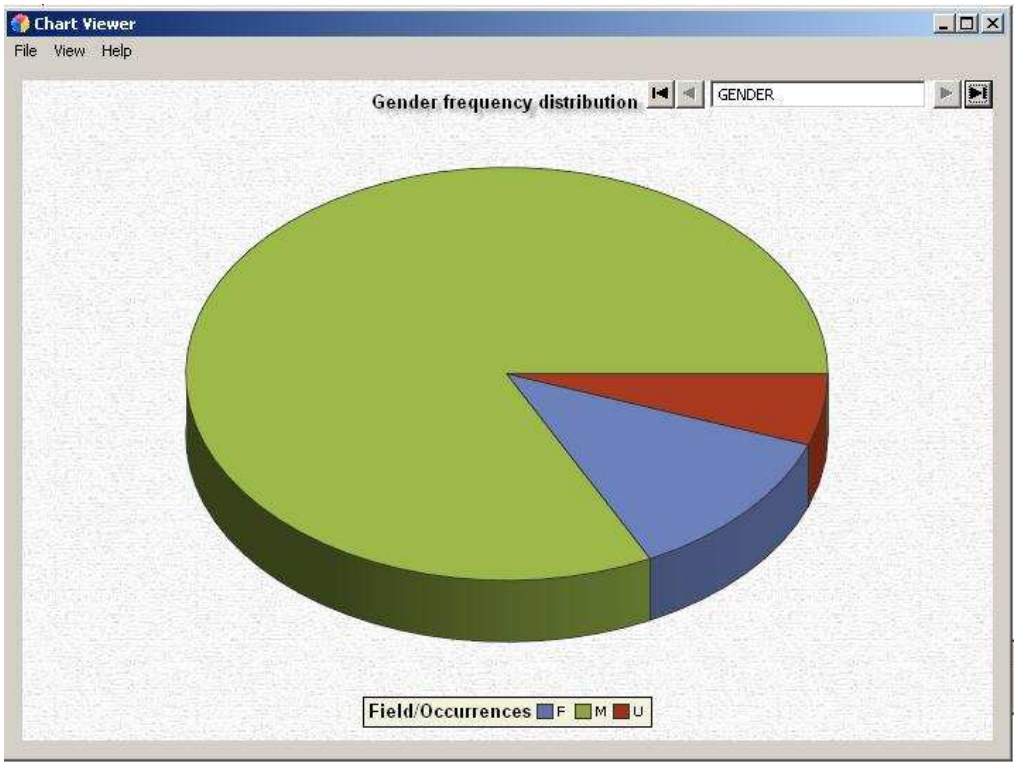
Tabulka 39-Určení pohlaví

Podrobnější analýza za pomoci databáze

Podrobnější analýzu uděláme pomocí sql scriptu gender.sql uloženým na CD.Výstupem této podrobnější analýzy je CSV soubor gender_final.csv, zobrazující každý záznam obohacený o pohlaví.

JMENO	PRIJMENI	GENDER
		U
Sybase Francie		U
	Posedníková	F
???	???	U
Jerden Van	+	U
Lukáš	Abazid	M
Jan	Abbrent	M
Ivan	Abdul	M

Nad tímto souborem ještě uděláme pomocí SAS DataFluxu frekvenční analýzu jejímž výstupem je následující koláčový graf zobrazující distribuci pohlaví ve zdrojových datech



Graf analýzy pohlaví

Výsledný graf ukazuje, že většina kontaktů je mužského pohlaví, což odpovídá i předpokladům, protože se jedná o kontakty z oblasti IT.

Vyhodnocení doplnění pohlaví

Abychom posoudili úspěšnost doplnění pohlaví ke každému kontaktu uděláme frekvenční analýzu jednotlivých hodnot.

Kombinace pohlaví zjištěných ze jména a příjmení	Výsledná hodnota
FF, MM	J – jistota
UM, MU, FU, UF	P – vysoká pravděpodobnost správnosti
FM, MF	S – spor mezi zjištěnými hodnotami
UU	N - nezjištěno

Tabulka 40-Vyhodnocení gender analýzy

Výstup frekvenční analýzy je uložen v html reportu Analýza doplnění pohlaví.htm.

Analýza doplnění pohlaví

Kombinace	Počet
J	2761
P	12603
S	82
U	829

Výstupní hodnoty frekvenční analýzy nám ukazují, že ve většině případů bylo pohlaví zjištěno pouze z jednoho dílčího údaje, tj. buď ze jména nebo příjmení. Je samozřejmé, že většinou to bylo ze jména, protože z příjmení, pokud se nejedná o příjmení typicky ženské tj. zakončeného na "ová" toho nejsme schopni moc zjistit.
V každém případě o pohlaví byla obohacena většina záznamů a existuje jen minimum záznamů, kde došlo ke sporu případně nebylo pohlaví určeno ani ze jména ani z příjmení.

Tabulka *Vyhodnocení obohacení dat* shrnuje úspěšnost obohacení dat.

	Metrika	Hodnota
Obohacení dat	Počet hodnot	16275
	Počet obohacených hodnot	15364

Tabulka 41-Vyhodnocení obohacení dat

.

6.5 Stanovení hodnot výsledných metrik

Nyní provedeme výsledné vyhodnocení všech dílčích i souhrnných metrik dat a rozebereme příčiny a důsledky hodnot těchto metrik.

6.5.1 Vyhodnocení oblasti Kvalita dat

Úplnost

Tabulka *Úplnost dat* obsahuje přehled hodnot dílčích metrik posuzujících úplnost jednotlivých sloupců. Je z ní zřejmé, že nejslabším článkem co se úplnosti týče je titul, který má ovšem ve výsledném vyhodnocení má váhu 0 a celkový výsledek tedy neovlivní.

Úplnost	Počet hodnot	Počet vyplněných hodnot	Procento vyplněných hodnot
Jméno	16275	15824	97,23%
Příjmení	16275	16270	99,97%
Titul	16275	4631	28,45%
Pozice	16275	10737	65,97%
Email	16275	14290	87,80%
Název společnosti	16275	16161	99,30%
Telefon	16275	13792	84,74%
Ulice	16275	16080	98,80%
Město	16275	16111	98,99%
Plat	16275	16271	99,98%

Tabulka 42-Úplnost dat

Tabulka *Vyhodnocení úplnosti dat* znázorňuje hodnoty metrik přepočítané vzhledem k jejich váze a celkové vyhodnocení souhrnné metriky úplnost.

Úplnost	Procento vyplněných hodnot	Váha metriky	Přepočet
Jméno	97,23%	0,1	9,72%
Příjmení	99,97%	0,4	39,99%
Titul	28,45%	0	0,00%
Pozice	65,97%	0,05	3,30%
Email	87,80%	0,1	8,78%
Název společnosti	99,30%	0,2	19,86%
Telefon	84,74%	0,05	4,24%
Ulice	98,80%	0,02	1,98%
Město	98,99%	0,03	2,97%
Plat	99,98%	0,05	5,00%
Výsledná hodnota metriky úplnost			**95,83%**

Tabulka 43-Vyhodnocení úplnosti dat

Správnost

Tabulka *Správnost dat* obsahuje přehled hodnot dílčích metrik posuzujících správnost jednotlivých sloupců. V tomto případě považujeme prázdné hodnoty za nesprávné s výjimkou titulu. Z tabulky je vidět, že příjmení a název společnosti, pro nás nejdůležitější, jsou relativně správné.

Správnost	Počet hodnot	Počet ověřených hodnot	Procento ověřených hodnot
Jméno	16275	15573	95,69%
Příjmení	16275	15102	92,79%
Titul	4631	4380	94,58%
Název společnosti	16275	15211	93,46%
Město	16275	13941	95,69%
Plat	16275	16259	99,90%

Tabulka 44-Správnost dat

Tabulka *Vyhodnocení správnosti dat* znázorňuje hodnoty dílčích metrik správnosti přepočítané vzhledem k jejich váze a celkové vyhodnocení souhrnné metriky správnost.

Správnost	Procento ověřených hodnot	Váha	Přepočet
Jméno	95,69%	0,2	19,14%
Příjmení	92,79%	0,4	37,12%
Titul	94,58%	0,05	4,73%
Název společnosti	93,46%	0,2	18,69%
Město	95,69%	0,05	4,78%
Plat	99,90%	0,1	9,99%
Výsledná hodnota metriky správnost			**94,45%**

Tabulka 45-Vyhodnocení správnosti dat

Formát

Tabulka *Formát dat* obsahuje přehled hodnot dílčích metrik posuzujících formát emailu a telefonu. Z tabulky je patrné, že problém je zejména se sloupcem telefon, do jisté míry je to dáno i značnou neúplností tohohle sloupce.

Formát	Počet hodnot	Počet ověřených hodnot	Procento ověřených hodnot
Email	16275	13941	85,66%
Telefon	16275	11914	73,20%

Tabulka 46-Formát dat

Tabulka *Vyhodnocení formátu dat* znázorňuje hodnoty dílčích metrik formátu přepočítané vzhledem k jejich váze a celkové vyhodnocení souhrnné metriky formát.

Formát	Procento ověřených hodnot	Váha	Přepočet
Email	85,66%	0,5	42,83%
Telefon	73,20%	0,5	36,60%
Výsledná hodnota metriky formát			**79,43%**

Tabulka 47-Vyhodnocení formátu dat

Konzistence

Tabulka *Vyhodnocení konzistence dat* shrnuje výsledky aplikace konzistence měst a ulic. Je zřejmé, že i přes relativně benevolentní podmínky, co se týče zápisu daných polí, není konzistence stoprocentní. Nicméně i tohle je jistě částečně důsledkem špatné úplnosti dat.

Konzistence	Počet hodnot	Počet konzistentních hodnot	Procento konzistentních hodnot
	16275	13160	**80,86%**

Tabulka 48-Vyhodnocení konzistence dat

6.5.2 Vyhodnocení oblasti Deduplikace

Tabulka *Vyhodnocení deduplikace dat* obsahuje výsledky měření unifikace dat. Je z ní zřejmé, že data jsou značně duplikovaná, duplikovaná až do té míry, že duplicity ovlivňují správnost dat, speciálně pokud pro daný záznam je několikrát zopakována nesprávná hodnota.

Deduplikace	Počet hodnot	Počet unifikovaných hodnot	Procento unifikovaných hodnot
	16275	12466	**76,60%**

Tabulka 49-Vyhodnocení deduplikace dat

6.5.3 Vyhodnocení oblasti Obohacení dat

Tabulka *Vyhodnocení obohacení* dat vyhodnocuje úspěšnost obohacení dat. Pokud bychom brali v úvahu pouze úplná data bylo by obohacení dat téměř stoprocentní. Na procesu obohacení dat není prozatím třeba nic zlepšovat.

Obohacení dat	Počet hodnot	Počet obohacených hodnot	Procento obohacených hodnot
	16275	15364	**94,40%**

Tabulka 50-Vyhodnocení obohacení dat

6.5.4 Výsledné vyhodnocení

Tabulka *Výsledné vyhodnocení* obsahuje celkový přehled vypočítaných souhrnných metrik pro jednotlivé oblasti. Je z ní zřejmé, že data nejsou v žádném případě ideální.

Nejslabším článkem a největším problémem dat je jejich značná duplikace, zde vidím největší prostor pro zlepšení. Další zlepšení kvality dat by bylo určitě zlepšení úplnosti dat, která negativně ovlivňuje nejenom správnost dat, protože prázdné hodnoty, se počítají do nesprávných, ale také konzistenci dat, formát dat a výsledek obohacování dat. Nejlépe bych hodnotila pravděpodobně výsledek obohacení dat, které je v rámci vyplněných údajů téměř stoprocentní.

Kvalita dat	**Procento hodnot**	**Váha**	**Přepočet**
Správnost (M1)	90,82%	0,4	36,33%
Úplnost (M2)	95,83%	0,4	38,33%
Formát(M3)	79,43%	0,1	7,94%
Konzistence (M4)	80,86%	0,1	8,09%
Výsledná hodnota oblasti Kvalita dat			90,69%

Tabulka 51-Vyhodnocení Kvality dat

Oblast	**Hodnota**
Kvalita dat Q1	**90,69%**
Deduplikace Q2	**76,60%**
Obohacení dat Q3	**94,40%**

Tabulka 52-Výsledné vyhodnocení

7 Závěr

Cílem této práce bylo shrnout problematiku datové kvality a problematiku měření kvality dat a navrhnout systém pro vyhodnocování stavu dat z hlediska jejich kvality (tzv. data profiling) a tento systém aplikovat na nějaká reálná data.
Na závěr lze tedy říci, že cíl práce podle zadání byl splněn. Vytvořený systém objektivně měří kvalitu dat, konkrétně v našem případě pro seznam kontaktů.
Za pomoci výsledků tohoto měření jsme schopni kvalitu dat posoudit a učinit nápravná opatření pro zvýšení kvality dat. Pravidelnou aplikací navrženého systému, lze sledovat trendy v kvalitě dat a periodicky provádět systematické zlepšování kvality dat na takovou úroveň, aby umožnila co možná nejlepší fungování business procesů dané společnosti.
Dlužno ovšem říci, že měření správnosti dat porovnáváním je také závislé na kvalitě dat registru se kterým porovnáváme a bohužel žádná data nejsou stoprocentně kvalitní, takže i naše měření je i díky tomu nepřesné. Stejně tak se málokdy podaří navrhnout ideální pravidla datové kvality, někdy je velice obtížné podchytit všechny souvislosti.
Dále je nutno připomenout, že kvalita dat je pojem relativní a ačkoli máme objektivní systém na vyhodnocení kvality dat, nemusí být tento systém vhodný za každé situace, vždy záleží na konkrétních business požadavcích.

Přínos práce

Hlavní přínos práce spatřuji v navržení obecného systému pro posuzování kvality dat. Pro libovolné seznamy kontaktů se dá tohoto systému využít téměř bez modifikací. S mírnými modifikacemi se potom tohoto systému dá využit pro jakákoliv jiná data.
Dalším důležitým přínosem práce je vytvoření rozhraní pro posuzování správnosti a konzistence dat. Pro české prostředí byly vytvořeny seznamy k porovnávání jmen, příjmení, titulů a měst. Pro posouzení konzistence dat byl získán seznam měst s korespondujícími ulicemi. Stejně tak bylo popsáno, jak lze tyhle seznamy získat a vytvořit, protože tyto údaje se často mění a co platí dnes, zítra už bude nedostatečné.

Možná zlepšení

Zlepšení by se mohla týkat dvou oblastí. První oblast je zlepšení samotného vyhodnocovacího systému, druhou možností naopak zlepšení a zkvalitnění pravidel pro výpočet jednotlivých metrik.
Co se týče vyhodnocovacího systému bylo by možné samozřejmě přidat do systému v souvislosti s posuzovanými daty další dimenze a metriky.
U výpočtu metrik se nabízí zlepšení jak z pohledu porovnávání dat, pro posouzení konzistence měst a ulic by se dal využít přesný registr pro existenci daného adresního místa a nejenom konzistence město-ulice.
Dalším zlepšením by bylo ověřování správnosti telefonních čísel za pomoci registru existujících telefonních čísel ČTÚ a nejenom strohá kontrola formátu. Obdobně u názvu firmy by se dalo přistupovat k datům z obchodního rejstříku a nejenom ze jména firmy podle pravidel usuzovat, zda se jedná o název firmy či nikoli.

8 Seznam použité literatury a internetových zdrojů

[1]	**Data quality assement** Leo L. Pipino, Yang W. Lee, and Richard Y. Wang
[2]	**A Hierarchical Approach to Improving Data Quality** Marcey L. Abate, Kathleen V. Diegert Sandia National Laboratories and Heather W. Allen, Heather Allen &Associates
[3]	**Data quality assesment (2007)** Arkady Maydanchik Technics Publications, LCC
[4]	**Wikipedia** http://cs.wikipedia.org/wiki/Titul
[5]	**Magic Quadrant for Data Quality Tools** http://www.gartner.com/technology/about/
[6]	**Ministerstvo vnitra** http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni.aspx
[7]	**Seznam PSČ** http://www.julda.cz/2009/05/psc-mest-a-obci-seznam-ke-stazeni/

9 Terminologický slovník

ČTÚ	Český telekomunikační úřad
Data	Reprezentace faktů o věcech nebo entitách v reálném světě
Data cleasing	Čištění dat, proces zlepšování datové kvality
Data profiling	Zkoumání dat dostupných v existujících zdrojích dat a vytváření statistik a informací o těchto datech
Data quality assesment	Měření kvality dat
DataFlux	SW datové kvality od firmy SAS
Datová kvalita	Schopnost dat odpovídat business požadavkům
Deduplikace	Odstranění duplicitních záznamů
dfPower Studio	Klientská část DataFluxu sloužící především pro tvorbu business pravidel
Dimenze	Formalizovaná perspektiva reality a nástroj pro monitorování hodnoty metriky v daném kontextu
Geocoding	Proces zpětného přiřazení GPS souřadnic na základě názvu města či ulice
Householding	Identifikace záznamů, které patří k nějaké skupině
Matching	Porovnání dat na základě nějakého pravidla
Metrika	Parametr, který může být číselně vyjádřen a vypočítán
Parsing	Rozpoznání sémantických prvků uložených v atributu a určení jednotlivých datových komponent
QKB	Znalostní báze obsahující různá schémata a definice
Standardizace	Proces transformace datových elementů do standardního formátu datového elementu
Unifikace	Identifikaci záznamů a seskupení těchto záznamů do skupin, které patří ke konkrétnímu subjektu

10 Příloha - přehled tabulek

Tabulka 1-Levely data profilingu ... 23
Tabulka 2-Definiční tabulka metriky ... 24
Tabulka 3-Definiční tabulka dimenze ... 25
Tabulka 4-Navržené metriky ... 33
Tabulka 5-Váhy úplnosti ... 34
Tabulka 6-Váhy správnosti ... 35
Tabulka 7-Váhy formátu ... 35
Tabulka 8-Definiční tabulka metriky Úplnost ... 35
Tabulka 9-Definiční tabulka metriky Správnost ... 36
Tabulka 10-Definiční tabulka metriky Formát ... 36
Tabulka 11-Váhy kvality dat ... 37
Tabulka 12-Definiční tabulka Kvality dat ... 37
Tabulka 13-Definiční tabulka Unifikace ... 38
Tabulka 14-Definiční tabulka Obohacení dat ... 38
Tabulka 15-Profiling úplnosti dat ... 39
Tabulka 16-Unikátnost dat ... 40
Tabulka 17-Podrobnější profiling ... 40
Tabulka 18-Zákaldní profiling platu ... 41
Tabulka 19-Percentily platu ... 41
Tabulka 20-Krajní hodnoty platu ... 42
Tabulka 21-Frekvenční analýza platu ... 42
Tabulka 22-Vyhodnocení platu ... 43
Tabulka 23-Vyhodnocení příjmení ... 46
Tabulka 24-Vyhodnocení jména ... 47
Tabulka 25-Ověření titulů ... 47
Tabulka 26-Titul bez standardizace ... 48
Tabulka 27-Standardní tvar titulu ... 48
Tabulka 28-Vyhodnocení titulu ... 49
Tabulka 29-Vyhodnocení pozice ... 49
Tabulka 30-Vyhodnocení parsingu emailu ... 51
Tabulka 31-Vyhodnocení emailu ... 51
Tabulka 32-Identifikace názvu společnosti ... 52
Tabulka 33-Vyhodnocení názvu společnosti ... 52
Tabulka 34-Vyhodnoceni telefonu ... 54
Tabulka 35-Vyhodnocení města ... 55
Tabulka 36-Vyhodnocení konzistence ... 58
Tabulka 37-Duplicity ... 59
Tabulka 38-Vyhodnocení deduplikace ... 59
Tabulka 39-Určení pohlaví ... 59
Tabulka 40-Vyhodnocení gender analýzy ... 61
Tabulka 41-Vyhodnocení obohacení dat ... 61
Tabulka 42-Úplnost dat ... 62
Tabulka 43-Vyhodnocení úplnosti dat ... 62
Tabulka 44-Správnost dat ... 63
Tabulka 45-Vyhodnocení správnosti dat ... 63
Tabulka 46-Formát dat ... 63
Tabulka 47-Vyhodnocení formátu dat ... 64
Tabulka 48-Vyhodnocení konzistence dat ... 64
Tabulka 49-Vyhodnocení deduplikace dat ... 64
Tabulka 50-Vyhodnocení obohacení dat ... 64
Tabulka 51-Vyhodnocení Kvality dat ... 65
Tabulka 52-Výsledné vyhodnocení ... 65

11 Příloha - přehled souborů na CD

Adresářová struktura :

- **Source Codes** – obsahuje zdrojové kódy
  Oracle – obsahuje pomocné skripty v pl/sql
  SAS – obsahuje architect, match a profile joby, vytvořená schémata
- **Zdroje –** obsahuje zdrojový soubor kontakty.txt a soubory s registry pro porovnávání dat
- **Výstupy** – adresář obsahuje podadresáře s jednotlivými reporty
- **Text** – obsahuje soubor data_profiling.doc s textem diplomové práce